Convergence Analysis of Riemannian Stochastic Approximation Schemes

Alain Durmus, Pablo Jiménez, Eric Moulines, Salem Said, Hoi-To Wai

Convergence Analysis of Riemannian Stochastic Approximation Schemes

Alain Durmus[1], Pablo Jiménez[2], Éric Moulines[2], Salem Said[3], and Hoi-To Wai[4]

[1]*Centre Borelli, UMR 9010, École Normale Supérieure Paris-Saclay, CNRS*
[2]*CMAP, UMR 7641, École Polytechnique, CNRS*
[3]*Laboratoire IMS, UMR 5218, CNRS, Université de Bordeaux*
[4]*Department of SEEM, The Chinese University of Hong Kong*

May 27, 2020

Abstract

This paper analyzes the convergence for a large class of Riemannian stochastic approximation (SA) schemes, which aim at tackling stochastic optimization problems. In particular, the recursions we study use either the exponential map of the considered manifold (geodesic schemes) or more general retraction functions (retraction schemes) used as a proxy for the exponential map. Such approximations are of great interest since they are low complexity alternatives to geodesic schemes. Under the assumption that the mean field of the SA is correlated with the gradient of a smooth Lyapunov function (possibly non-convex), we show that the above Riemannian SA schemes find an $\mathcal{O}(b_\infty + \log n/\sqrt{n})$-stationary point (in expectation) within $\mathcal{O}(n)$ iterations, where $b_\infty \geq 0$ is the asymptotic bias. Compared to previous works, the conditions we derive are considerably milder. First, all our analysis are global as we do not assume iterates to be a-priori bounded. Second, we study biased SA schemes. To be more specific, we consider the case where the mean-field function can only be estimated up to a small bias, and/or the case in which the samples are drawn from a controlled Markov chain. Third, the conditions on retractions required to ensure convergence of the related SA schemes are weak and hold for well-known examples. We illustrate our results on three machine learning problems.

Email: alain.durmus@cmla.ens-cachan.fr,pablo.jimenez-moreno@polytechnique.edu
eric.moulines@polytechnique.edu,salem.said@u-bordeaux.fr,htwai@se.cuhk.edu.hk

1 Introduction

This paper is concerned with the root finding problem on a smooth Riemannian manifold Θ:

$$\text{find} \quad \theta \in \Theta \quad \text{satisfying} \quad h(\theta) = 0_\theta, \quad \text{where} \quad h(\theta) = \int_{\mathsf{X}} H_\theta(x) \mathrm{d}\pi_\theta(x) \;, \tag{1}$$

such that $h : \Theta \to \mathrm{T}\Theta$ is a smooth vector field, called the *mean vector field*, for any $\theta \in \Theta$, π_θ is a distribution over $(\mathsf{X}, \mathcal{X})$ and $H : \Theta \times \mathsf{X} \to \mathrm{T}\Theta$ is a bimeasurable function. In particular, note that this framework includes stochastic optimization problems where the objective function $f : \Theta \to \mathbb{R}^*$ is smooth (but possibly non-convex even in the geodesic sense), taking $h = \operatorname{grad} f$, where grad is the Riemannian gradient operator, see Appendix A.8. Problem (1) is motivated by applications in principal component analysis, combinatorial optimization, and geometric barycenter.

Our aim is to study the stochastic approximation (SA) schemes adapted to a Riemannian setting to compute an *approximate stationary point* of (1). In the case where the exponential map on Θ can be evaluated, so geodesic curves can be explicitly calculated, the Riemannian SA scheme that we consider is a natural adaptation of the standard Robbins-Monro algorithm [33], namely the iterates are updated via the exponential map applied to Monte Carlo estimates of $h(\theta)$ for $\theta \in \Theta$. Furthermore, the present paper considers the generalizations of the above methodology where (a) the considered estimators are potentially *biased*, *i.e.* they are either based on a Markov chain targeting π_θ and more generally the right-hand side of (1) is not satisfied *i.e.* $\int_{\mathsf{X}} H_\theta(x)\mathrm{d}\pi_\theta(x) = h(\theta) + b(\theta)$, where $b : \Theta \to \mathrm{T}\Theta$ is a smooth vector field modeling the bias introduced by the scheme; (b) the updates are computed using *retraction* functions which are computationally cheaper to evaluate than the exponential map but provide good numerical approximation of the latter. These generalizations are important as we adapt the Riemannian SA scheme to instances of (1) with various constraints in data acquisitions and computation.

A lot of effort has been paid in the analysis of SA in the last decade due to its applications in machine learning and signal processing. As mentioned previously, SA encompasses stochastic gradient (SG) methods but is not limited to the case of $h = \operatorname{grad} f$. Most common examples for which the latter condition is not satisfied are second order methods, online Expectation Maximization algorithms [12], Q-learning [23] or policy gradient [7]. In this sense, our work is in line with recent studies of SA in the Euclidean setting under mild conditions on the mean-field h and which can be applied to stochastic optimization seeking at minimizing non-convex objective function, see e.g. [19, 25] and the references therein. In addition, we also extend results regarding biased SA [37, 25] to the Riemannian setting.

SA on Riemannian manifolds has attracted attentions recently. The pioneering work [9] proved the asymptotic convergence of Riemannian SA to a critical point using martingale techniques usually used in the analysis of SA in Euclidean spaces, see [8, 15]. Under the assumption that the objective function f is geodesically convex (g-convex), [43] showed that the Riemannian SGD scheme finds an ϵ-optimal solution in $\mathcal{O}(1/\epsilon^2)$ iterations. This result has been improved since then in the authors' follow up works [42, 44]. However, note that these papers focus on geodesic schemes which can be computationally expensive. Another related work is [40] which considers a *retraction* based averaging scheme, and improves the convergence rate to $\mathcal{O}(1/\epsilon)$ by assuming f is geodesically strongly convex and the SA iterates stay in a compact set. In addition, this paper also considers non-convex settings and shows a central limit theorem under strong assumptions for a sequence $(\psi(\overline{\theta}_n))_{n\in\mathbb{N}}$ where $(\overline{\theta}_n)_{n\in\mathbb{N}}$ is an appropriate averaged sequence based on $(\theta_n)_{n\in\mathbb{N}}$ and for some function $\psi : \Theta \to \mathrm{T}_{\theta_\star}\Theta$, $\theta_\star$ is the minimum of f. To deal with relatively general conditions on the mean-field h, [35] analyzes stochastic recursion schemes adapting the well-known ODE method [8, 26] to the Riemannian setting to obtain asymptotic convergence. Note however that this study is limited to the case where Θ is compact. Analysis of Riemannian SGD for non-convex objective has been addressed in [22] which provides a similar result to one of ours in the case of unbiased Riemaniann SA. However, they require a strong assumption on the retraction map, the objective function and the manifold. We should also mention that a few other papers have studied deterministic optimization on Riemannian manifold and the convergence to local minimum in non-convex settings, see e.g. [10, 38, 13].

Our contributions are three-fold. (A) In Section 3, we perform a global convergence analysis of a *biased* geodesic Riemannian SA scheme, without assuming a bounded domain for the iterates $(\theta_n)_{n\in\mathbb{N}}$ nor strong g-convexity type of assumptions. Under these relaxed settings, the considered geodesic SA scheme finds an $\mathcal{O}(b_\infty + \log n/\sqrt{n})$-stationary point in $\mathcal{O}(n)$ iterations in expectation, if the gradient samples are asymptotically $\mathcal{O}(b_\infty)$-biased. We cover settings when the noise is a sequence of martingale increments or stems from a Markovian dynamics, see Theorem 2 & 3. (B) In Section 4, we consider retraction schemes with first or second order retraction. Similar to the geodesic schemes, the methods we introduce find an $\mathcal{O}(b_\infty + \log n/\sqrt{n})$-stationary point in $\mathcal{O}(n)$ iterations in expectation, see Theorem 5 & 6. The conditions we consider are stronger than in the case of geodesic schemes but relax the ones required in previous works. Further, we illustrate on several examples that the required conditions on the retraction function hold. (C) In Section 5, we consider three example applications and show that the required convergence conditions are satisfied. These applications are novel as they cannot be handled using previous analysis.

Notations For any two sequences of real numbers $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$, we write $u_n = \mathcal{O}(v_n)$ when there exist $n_0 \in \mathbb{N}$ and $M_0 \in \mathbb{R}_+$ such that for any $n \geq n_0$, $|u_n| \leq M_0 v_n$. If $u_n = \mathcal{O}(v_n)$ and $v_n = \mathcal{O}(u_n)$, we write $u_n = \Omega(v_n)$. We denote the tangent space of Θ at θ by $\mathrm{T}_\theta\Theta$ and its tangent bundle $\mathrm{T}\Theta$. If Θ_0 and Θ_1 are two smooth manifolds, for any smooth function $f : \Theta_0 \to \Theta_1$, we denote its derivative by $\mathrm{D}f : \mathrm{T}\Theta_0 \to \mathrm{T}\Theta_1$. The Riemannian metric on Θ is denoted by g but for ease of notation and if there is no risk of confusion, for any $\theta \in \Theta$, $u, v \in \mathrm{T}_\theta\Theta$, we should denote $\mathrm{g}_\theta(u,v) = \langle u, v\rangle_\theta$ and $\|u\|_\theta^2 = \mathrm{g}_\theta(u,u)$. $\mathrm{T}^{\gamma}_{t_0 t_1} : \mathrm{T}_{\gamma(t_0)}\Theta \to \mathrm{T}_{\gamma(t_1)}\Theta$ stands for the parallel transport associated to the Levi-Civita connection along a curve $\gamma : \mathrm{I} \to \Theta$ from $\gamma(t_0)$ to $\gamma(t_1)$. In the interest of space, we leave detailed definitions and generalities on Riemannian geometry to Appendix A.

2 Riemannian Stochastic Approximation Schemes

Let $(X_n)_{n\in\mathbb{N}^*}$ be a stochastic process defined on the filtered probability space $(\mathsf{X}, \mathcal{X}, \mathbb{P}, (\mathcal{F}_n)_{n\in\mathbb{N}})$, $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted. The function $\mathrm{Ret} : \mathrm{T}\Theta \to \Theta$ is a retraction and for any $\theta \in \Theta$, Ret_θ stands for the restriction of Ret to $\mathrm{T}_\theta\Theta$. We consider the cases where the function Ret can be either the exponential map or a computationally efficient proxy. The present paper studies stochastic approximation (SA) sequences $(\theta_n)_{n\in\mathbb{N}}$, used as approximate solutions of (1), starting from $\theta_0 \in \Theta$ and defined by the recursion

$$\theta_{n+1} = \mathrm{Ret}_{\theta_n}\left\{\eta_{n+1}\left(H_{\theta_n}(X_{n+1}) + b_{\theta_n}(X_{n+1})\right)\right\} , \tag{2}$$

where $H, b : \Theta \times \mathsf{X} \to \mathrm{T}\Theta$ are bi-measurable functions, and $(\eta_n)_{n\in\mathbb{N}^*}$ is a sequence of positive stepsizes. For any $n \in \mathbb{N}^*$, $H_{\theta_n}(X_{n+1})$ is a noisy version of $h(\theta_n)$ and $b_{\theta_n}(X_{n+1})$ is an additional bias term. Eq. (2) describes the recursion where θ_n is moved along the direction given by $\eta_{n+1}(H_{\theta_n}(X_{n+1}) + b_{\theta_n}(X_{n+1}))$ on the tangent space $\mathrm{T}_{\theta_n}\Theta$ originated at θ_n, and the retraction Ret_{θ_n} "projects" the updated iterate back into Θ. In this sense, the vector $H_{\theta_n}(X_{n+1}) + b_{\theta_n}(X_{n+1}) \in \mathrm{T}_{\theta_n}\Theta$ can be interpreted as the stochastic update vector in Euclidean SA.

Let us discuss some basic assumptions to be used throughout this paper. Consider the following condition on the Riemannian manifold.

A 1. *Θ is a geodesically complete Riemannian manifold of dimension $d \in \mathbb{N}^*$, i.e. the exponential function* Exp *is defined over* $\mathrm{T}\Theta$.

For a definition of the Riemannian exponential mapping, see Appendix A.5. We assume in addition the existence of a Lyapunov function for the mean vector field h.

A 2. *There exists a continuously differentiable function $V : \Theta \to \mathbb{R}^*$ satisfying the conditions below.*
(a) There exist constants $\underline{c}, \bar{c} > 0$ such that, for any $\theta \in \Theta$,

$$\underline{c}\,\|h(\theta)\|_\theta^2 \le -\langle \operatorname{grad} V(\theta), h(\theta)\rangle_\theta \ , \qquad \|\operatorname{grad} V(\theta)\|_\theta \le \bar{c}\,\|h(\theta)\|_\theta \ .$$

(b) The Riemannian gradient $\operatorname{grad} V$ *is geodesically L-Lipschitz, i.e. there exists $L \ge 0$ such that for any $\theta_0, \theta_1 \in \Theta$, and geodesic curve $\gamma : [0,1] \to \Theta$ between θ_0 and θ_1,*

$$\|\operatorname{grad} V(\theta_1) - \mathrm{T}_{01}^{\gamma}\operatorname{grad} V(\theta_0)\|_{\theta_1} \le L\ell(\gamma) \ , \tag{3}$$

where $\ell(\gamma) = \|\dot{\gamma}(0)\|_{\theta_0}$ is the length of the geodesic.

The first inequality in (a) strengthens the Lyapunov's second method for stability and Lasalle's invariance principle for integral curves or ordinary differential equations; see [27]. In fact, it implies $\|h(\theta)\|_\theta \le \tilde{c}\,\|\operatorname{grad} V(\theta)\|_\theta$ for any $\theta \in \Theta$, for $\tilde{c} = \underline{c}^{-1} > 0$. Meanwhile, (b) is satisfied if V has a continuous Riemannian Hessian $\operatorname{Hess} V(\theta) : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$ with a bounded operator norm for all $\theta \in \Theta$, see Lemma 10 in Appendix A.9. Overall, conditions (a) and (b) ensure the stability of the recursion (2) as they imply that $h(\theta)$ is sublinear, *i.e.*, for any $\theta_0, \theta_1 \in \Theta$, there exists $C \geq 0$ such that for any geodesic curve $\gamma : [0,1] \to \Theta$ between θ_0, θ_1, one has $\|h(\theta)\|_\theta \leq C(\ell(\gamma)+1)$. Importantly, **A**1, **A**2 allow us to generalize the well-known descent lemma to the Riemannian setting, as follows:

Lemma 1. *Assume* ***A****1,* ***A****2-(b) hold. For any $\theta_0, \theta_1 \in \Theta$ and geodesic curve $\gamma : [0,1] \to \Theta$ between θ_0, θ_1,*

$$\left| V(\theta_1) - V(\theta_0) - \langle \operatorname{grad} V(\theta_0), \dot{\gamma}_0 \rangle_{\theta_0} \right| \leq L\ell(\gamma)^2/2 \, .$$

The proof is postponed to Appendix B. This result was stated in [43] without proof as a consequence of **A**2-(a). Moreover, compared to [2, Lemma 7.4.7], our result holds for any geodesic curves, and is not limited to the length-minimizing ones.

Finally, we assume that the bias is uniformly bounded.

A3. *There exists a constant b_∞ such that $\sup_{(\theta,x)\in\Theta\times\mathsf{X}} \|b_\theta(x)\|_\theta \leq b_\infty$.*

The present paper provides *a priori* non-asymptotic guarantees for the SA scheme (2) to find an *(approximate) stationary point* where $\|h(\theta)\|_\theta^2 \approx \widetilde{c} b_\infty$ for some $\widetilde{c} > 0$. Roughly, a non-asymptotic performance guarantee ensures the ability of the scheme (2) to produce an iterate $\theta_{I_n} \in \{\theta_0, \ldots, \theta_n\}$ such that $\mathbb{E}[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2] \leq \epsilon + \widetilde{c} b_\infty$, for given $n \in \mathbb{N}^*$ and $\epsilon > 0$, where for any $k \in \mathbb{N}^*$, we define the randomized stopping rule I_k independent of $(\theta_0, X_1, X_2, ...)$ with distribution for any $\ell \in \{0, \ldots, k\}$,

$$\mathbb{P}\left(I_k = \ell\right) = \left(\textstyle\sum_{i=0}^k \eta_{i+1} \right)^{-1} \eta_{\ell+1} \, . \tag{4}$$

To simplify notations in our subsequent discussions, define for any $n \in \mathbb{N}$ and $p \geq 1$,

$$\Gamma_n^{(p)} = \textstyle\sum_{k=1}^n \eta_k^p \, , \qquad \Gamma_n = \Gamma_n^{(1)} \, .$$

For any $x \in \mathsf{X}$ and $\theta \in \Theta$, we define the error in estimating the mean vector field as:

$$e_\theta(x) = H_\theta(x) - h(\theta) \, . \tag{5}$$

3 Analysis of Geodesic SA Schemes

In this section, the retraction Ret appearing in (2) is the Riemannian exponential map of Θ, Ret $=$ Exp. Such scheme will also be referred to as the *geodesic SA scheme*.

Consider first the martingale setting formalized by the following conditions.

MD1. *The sequence $(e_{\theta_n}(X_{n+1}))_{n\in\mathbb{N}}$ is a martingale difference sequence with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, i.e. for any $n \in \mathbb{N}^*$, $\mathbb{E}\left[e_{\theta_n}(X_{n+1})|\mathcal{F}_n\right] = 0$. Also, there exist $\sigma_0^2, \sigma_1^2 < \infty$ such that for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}\left[\|e_{\theta_n}(X_{n+1})\|_{\theta_n}^2 \middle| \mathcal{F}_n\right] \leq \sigma_0^2 + \sigma_1^2 \, \|h(\theta_n)\|_{\theta_n}^2 \, .$$

We observe that:

Theorem 2. *Assume **A** 1-**A** 2-**A** 3-**MD** 1 hold. Consider $(\theta_n)_{n\in\mathbb{N}}$ defined by (2) with* $\mathrm{Ret} = \mathrm{Exp}$. *If $\sup_{k\in\mathbb{N}^*} \eta_k \leq \underline{c}/(6L(1+\sigma_1^2))$, then for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}\left[\|h(\theta_{I_n})\|^2_{\theta_{I_n}}\right] \leq 4\left(\underline{c}\Gamma_{n+1}\right)^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + (3L/2)(\sigma_0^2 + b_\infty)\Gamma^{(2)}_{n+1}\right\} + b_\infty\overline{c}/\underline{c}^2\ , \tag{6}$$

where $I_n \in \{0, ..., n\}$ is a discrete random variable with distribution defined by (4).

The proof is postponed to Appendix C.1.

A popular choice for the step size is $\eta_k = \rho/\sqrt{k+k_0}$ for some $\rho, k_0 > 0$. In this case, we have $\Gamma_{n+1} = \Omega(\sqrt{n})$ and $\Gamma^{(2)}_{n+1} = \mathcal{O}(\log n)$. Consequently, Theorem 2 shows that (2) finds an $\mathcal{O}(b_\infty + \log(n)/\sqrt{n})$-stationary point to (1) in $\mathcal{O}(n)$ iterations.

We now turn to the Markovian setting which is formalized by the following assumptions on the sequence $(X_n)_{n\in\mathbb{N}^*}$ and the stochastic vector field H.

MA 1. *There exists a Markov kernel P on $(\Theta \times \mathsf{X}) \times \mathcal{X}$ such that for any $n \in \mathbb{N}$ and bounded and measurable function $\varphi : \mathsf{X} \to \mathbb{R}_+$, $\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] = \int_{\mathsf{X}} \varphi(y) P_{\theta_n}(X_n, \mathrm{d}y)$. In addition, P satisfies the following conditions.*
(i) For any $\theta \in \Theta$, P_θ admits a unique invariant distribution π_θ satisfying $h(\theta) = \int_{\mathsf{X}} H_\theta(y)\mathrm{d}\pi_\theta(y)$.
(ii) Consider the error function $e_\theta(x)$ defined in (5). There exists a measurable function $\hat{e} : \Theta \times \mathsf{X} \to \mathrm{T}\Theta$ satisfying for any $x \in \mathsf{X}$, $\theta \in \Theta$, $\hat{e}_\theta(x) \in \mathrm{T}_\theta\Theta$ and

$$\hat{e}_\theta(x) - \int_{\mathsf{X}} P_\theta(x, \mathrm{d}y)\hat{e}_\theta(y) = e_\theta(x)\ . \tag{7}$$

(iii) There exist $e_\infty, \hat{e}_\infty \geq 0$ such that $\sup_{\theta\in\Theta, x\in\mathsf{X}} \|e_\theta(x)\|_\theta \leq e_\infty$, $\sup_{\theta\in\Theta, x\in\mathsf{X}} \|\hat{e}_\theta(x)\|_\theta \leq \hat{e}_\infty$.
(iv) There exists $L_{\hat{e}} \geq 0$ such that for any $\theta_0, \theta_1 \in \Theta$, and geodesic curve $\gamma : [0,1] \to \Theta$ between θ_0, θ_1,

$$\sup_{x\in\mathsf{X}} \left\| \int_{\mathsf{X}} P_{\theta_1}(x, \mathrm{d}y)\hat{e}_{\theta_1}(x) - \mathrm{T}^\gamma_{01}\left[\int_{\mathsf{X}} P_{\theta_0}(x, \mathrm{d}y)\hat{e}_{\theta_0}(y)\right]\right\|_{\theta_1} \leq L_{\hat{e}}\ell(\gamma)\ .$$

Condition (ii) assumes the existence of solutions to the Poisson equation (7), which can in general be established if the Markov kernel P_θ is V-geometrically or uniformly ergodic for any $\theta \in \Theta$, see [14, Section 21.2]. In addition, condition (iii) requires that function e is uniformly bounded and the corresponding solution to the Poisson equation (for fixed θ) as well. The uniform boundedness condition on $\hat{e}$ holds if, for example, the function e is uniformly bounded and for any $\theta \in \Theta$, P_θ is uniformly ergodic with convergence rate independent of θ. The latter could be relaxed using appropriate Lyapunov conditions following [17]. Finally, the last assumption (iv) is implied by smoothness conditions on the Markov kernel with respect to the SA parameter θ.

Under **A**1-**A**2-**A**3-**MA**1 and for a sequence of stepsize $(\eta_k)_{k\in\mathbb{N}^*}$, consider the constants

$$
\begin{aligned}
&D_{\hat{e}} = 1 + e_\infty + b_\infty + \hat{e}_\infty(L + \bar{c}a_2) \;, \qquad C(\eta_1) = \bar{c}\hat{e}_\infty(\eta_1 + 2) \;, \\
&C_{\hat{e}} = L_{\hat{e}}\bar{c}(e_\infty + b_\infty) + L\hat{e}_\infty(e_\infty + b_\infty + 1) + (3L/2)(b_\infty^2 + e_\infty^2) \;.
\end{aligned} \tag{8}
$$

We observe that:

Theorem 3. *Assume **A**1-**A**2-**A**3-**MA**1 hold. Let $(\eta_k)_{k\in\mathbb{N}^*}$ be a sequence of stepsizes and $a_1, a_2 \geq 0$ satisfying*

$$
\begin{aligned}
&\sup_{k\in\mathbb{N}^*}\{\eta_{k+1}/\eta_k\} \leq 1 \;, \qquad \sup_{k\in\mathbb{N}^*}\{\eta_k/\eta_{k+1}\} \leq a_1 \;, \\
&\sup_{k\in\mathbb{N}^*}\{|\eta_k - \eta_{k+1}|/\eta_k^2\} \leq a_2 \;, \qquad \sup_{k\in\mathbb{N}^*}\eta_k \leq c_1/(4(3L/2 + D_{\hat{e}})) \;.
\end{aligned} \tag{9}
$$

Consider $(\theta_n)_{n\in\mathbb{N}}$ defined by (2) with $\mathrm{Ret} = \mathrm{Exp}$. Then for any $n \in \mathbb{N}^*$,

$$
\mathbb{E}\left[\|h(\theta_{I_n})\|^2_{\theta_{I_n}}\right] \leq 4\left(\underline{c}\Gamma_{n+1}\right)^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + C(\eta_1) + C_{\hat{e}}\Gamma^{(2)}_{n+1}\right\} + b_\infty\bar{c}/\underline{c}^2 \;. \tag{10}
$$

where I_n has distribution defined by (4).

The proof is postponed to Appendix C.2.

We observe that the right hand side of (10) is akin to that of (6). Similar to the martingale noise setting, choosing the step size $\eta_k = \rho/\sqrt{k+k_0}$ and using (10) shows that (2) finds an $\mathcal{O}(b_\infty + \log(n)/\sqrt{n})$-stationary point to (1) in $\mathcal{O}(n)$ iterations.

4 Analysis of General Retraction SA Schemes

The previous section focused on the geodesic schemes that require performing the Riemannian exponential map Exp at each iteration. Evaluating the Exp map is often computationally hard as it requires computing complex functions such as matrix exponential. A popular idea is to use an efficient retraction function Ret that approximates Exp. One basic condition required on Ret is the following *local-rigidity* condition:

R 1. *For any $\theta \in \Theta$, $\mathrm{Ret}_\theta(0_\theta) = \theta$, where 0_θ is the zero element of $\mathrm{T}_\theta\Theta$ as vector space, and $\mathrm{DRet}_\theta(0_\theta) = \mathrm{Id}$.*

The following ensures that the inverse exponential map is defined on $\mathrm{Ret}_\theta(\mathrm{T}_\theta\Theta)$.

R 2. *For any $(\theta, u) \in \mathrm{T}\Theta$, $\mathrm{Ret}_\theta(u) \notin \mathrm{Cut}(\theta)$, where $\mathrm{Cut}(\theta) \subset \Theta$ is the cut locus of θ (see the definition in A.5).*

Under **R**2, for any $\theta \in \Theta$, the following function is well defined:

$$
\Phi_\theta = \mathrm{Exp}_\theta^{-1} \circ \mathrm{Ret}_\theta : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta \;, \tag{11}
$$

which defines a bundle map $\Phi : \mathrm{T}\Theta \to \mathrm{T}\Theta$. Using Φ_θ, the SA scheme (2) may be written as

$$\theta_{n+1} = \mathrm{Exp}_{\theta_n} \left\{ \eta_{n+1} \left(H_{\theta_n}(X_{n+1}) + b_{\theta_n}(X_{n+1}) + \Delta_{\theta_n, \eta_{n+1}}(X_{n+1}) \right) \right\} , \tag{12}$$

where $\Delta_{\theta_n, \eta_{n+1}}(X_{n+1})$ is the 'retraction bias' defined for any $(\theta, x) \in \Theta \times \mathsf{X}$ and $\eta > 0$ by

$$\Delta_{\theta,\eta}(x) = \eta^{-1} \Phi_\theta(\eta \{H_\theta(x) + b_\theta(x)\}) - \{H_\theta(x) + b_\theta(x)\} .$$

In the special case when $\mathrm{Ret} \equiv \mathrm{Exp}$, we have $\Phi_{\theta_n} \equiv \mathrm{Id}$ and thus $\Delta_{\theta,\eta}(x) = 0$ for any $(\theta, x) \in \Theta \times \mathsf{X}$ and $\eta > 0$. To control the retraction bias, we study the following conditions on Ret:

R 3. *Θ is a homogeneous Riemannian manifold with isometry group G (see the definition in A.7). For any $g \in \mathsf{G}$, $(\theta, u) \in \mathrm{T}\Theta$,*

$$g \cdot \mathrm{Ret}_\theta(u) \; = \; \mathrm{Ret}_{g \cdot \theta}(g \cdot u) \, . \tag{13}$$

In **R** 3, for an isometry $g : \Theta \to \Theta$, $g \cdot \theta = g(\theta)$ and $g \cdot u = \mathrm{D}g_\theta(u)$ is a vector in $\mathrm{T}_{g \cdot u}\Theta$. Example of retractions defined on special matrix manifolds satisfying **R**3 is given below. Then, we consider the following set of assumptions which lead to the definitions of regular *first-order* and *second-order* retractions [3]. These conditions will ensure that the first terms in the Taylor expansion of Φ_θ for $\theta \in \Theta$ do vanish.

R4. *For any $(\theta, u) \in \mathrm{T}\Theta$, there exists $\mathscr{L}^{(1)}(\theta) \geq 0$, such that $\sup_{t \in [0,1]} \left\| \mathrm{D}^2 \Phi_\theta(tu)[u, u] \right\|_\theta \leq \mathscr{L}^{(1)}(\theta) \left\| u \right\|_\theta^2$ where the function $\Phi_\theta : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$ is defined by (11).*

Let ∇ be the Levi-Civita connection of the metric g on Θ (see Appendix A.2). We consider:

R 5. *For any $(\theta, u) \in \mathrm{T}\Theta$, the following hold.*

(i) Setting $\gamma(t) = \mathrm{Ret}_\theta(tu)$ for $t \in \mathbb{R}$, the initial acceleration of the curve γ satisfies $\mathrm{D}_t \dot{\gamma}(0) = 0_\theta$, where D_t stands for the covariant derivative along γ (see [28, Theorem 4.24]).
(ii) There exists $\mathscr{L}^{(2)}(\theta) \geq 0$, such that $\sup_{t \in [0,1]} \left\| \mathrm{D}^3 \Phi_\theta(tu)[u, u, u] \right\|_\theta \leq \mathscr{L}^{(2)}(\theta) \left\| u \right\|_\theta^3$ where the function $\Phi_\theta : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$ is defined by (11).

A retraction Ret satisfying **R** 4 (respectively **R** 5) is called a regular first-order (respectively second-order) retraction. Note that **R** 5 does not imply **R** 4 nor vice versa. Intuitively, a first-order retraction approximates the Riemannian exponential map only to the first order, while a second-order retraction approximates the latter to the second order. These facts are formally established in the following result.

Lemma 4. *Assume **R** 1, **R** 2, **R** 3.*

(a) Under **R**4*, there exists* $\mathscr{L}_\infty^{(1)} \geq 0$ *such that* $\|\Phi_\theta(u) - u\|_\theta \leq \mathscr{L}_\infty^{(1)} \|u\|_\theta^2$ *for any* $(\theta, u) \in \mathrm{T}\Theta$*.*
(b) Under **R**5*, there exists* $\mathscr{L}_\infty^{(2)} \geq 0$ *such that* $\|\Phi_\theta(u) - u\|_\theta \leq \mathscr{L}_\infty^{(2)} \|u\|_\theta^3$ *for any* $(\theta, u) \in \mathrm{T}\Theta$*.*

The proof is postponed to Appendix D.1.

Lemma 4 bounds the retraction bias. With this lemma in mind, our strategy for analyzing (2) with a general retraction is to incorporate the retraction bias in the analysis done for the geodesic scheme. Before doing so, we strengthen **A**2 and **MD**1. Let $a > 0$.

A 4. *There exist* $V_\infty, h_\infty \geq 0$ *such that* $\sup_{\theta\in\Theta} \|\mathrm{grad}\, V(\theta)\|_\theta \leq V_\infty$ *and* $\sup_{\theta\in\Theta} \|h(\theta)\|_\theta \leq h_\infty$*.*

MD2 (a)**.** *Consider the noise function given by* (5)*. There exists* $\sigma_{(a)} \geq 0$ *such that almost surely, it holds that*

$$\mathbb{E}\left[\|e_{\theta_n}(X_{n+1})\|_{\theta_n}^a \,|\mathcal{F}_n\right] \leq \sigma_{(a)} \ ,$$

If **MD**2(a) holds for $a \in \mathbb{N}^*$, then **MD**2$(\tilde{a})$ holds for any $\tilde{a} \in \{0, 1, ..., a\}$ and $\sigma_{(\tilde{a})}$ will then stand for a constant such that almost surely, $\mathbb{E}[\|e_{\theta_n}(X_{n+1})\|_{\theta_n}^{\tilde{a}} |\mathcal{F}_n] \leq \sigma_{(\tilde{a})}$.

Depending on the order of retraction used and moment bound on the sequence $(e_{\theta_n}(X_{n+1}))_{n\in\mathbb{N}}$, we get the following convergence result for $(\theta_n)_{n\in\mathbb{N}}$.

Theorem 5. *Assume* **A**1*-***A**2*-***A**3*-***A**4*,* **R**1*-***R**2*-***R**3*,* **MD**1 *hold. Consider* $(\theta_n)_{n\in\mathbb{N}}$ *defined by* (2)*. For any* $n \in \mathbb{N}^*$ *and let* $I_n \in \{0, ..., n\}$ *be distributed according to* (4)*.*
(a) Assume **R**4*,* **MD**2(4)*. If* $\sup_{k\in\mathbb{N}^*} \eta_k \leq (4L + 6\mathscr{L}_\infty^{(1)} V_\infty)^{-1} \underline{c}$ *, then*

$$\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2\right] \leq 2(\underline{c}\Gamma_{n+1})^{-1} \left\{\mathbb{E}\left[V(\theta_0)\right] + c_0 \Gamma_{n+1}^{(2)} + c_1 \Gamma_{n+1}^{(4)}\right\} + 2b_\infty V_\infty / \underline{c} \ ,$$

where $c_0 = (3\mathscr{L}_\infty^{(1)} V_\infty + 2L)\{\sigma_{(2)} + b_\infty^2\}$*,* $c_1 = 54L(\mathscr{L}_\infty^{(1)})^2 \{h_\infty^4 + b_\infty^4 + \sigma_{(4)}\}$*.*
(b) Assume **R**5*,* **MD**2(6)*. If* $\sup_{k\in\mathbb{N}^*} \eta_k \leq (4L)^{-1} \underline{c}$ *, then*

$$\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2\right] \leq 2(\underline{c}\Gamma_{n+1})^{-1} \left\{\mathbb{E}\left[V(\theta_0)\right] + c_0 \Gamma_{n+1}^{(2)} + c_1 \Gamma_{n+1}^{(3)} + c_2 \Gamma_{n+1}^{(6)}\right\} + 2b_\infty V_\infty / \underline{c} \ ,$$

where $c_0 = 2L\{b_\infty^2 + \sigma_{(2)}\}$*,* $c_1 = 9\mathscr{L}_\infty^{(2)} V_\infty \{h_\infty^3 + b_\infty^3 + \sigma_{(3)}\}$*,* $c_2 = 486L(\mathscr{L}_\infty^{(2)})^2 \{h_\infty^6 + b_\infty^6 + \sigma_{(6)}\}$*.*

The proof is postponed to Appendix D.2.

For first-order retractions, one can relax the assumption of a bounded *fourth-order moment* for the noise term of sub-case (a), for a bounded first-order derivative for Φ_θ and a bounded *second-order moment* for the noise term instead, see Theorem 14 in the Appendix. For second-order retractions (b), the effects due to retraction are absorbed into the high-order terms $\Gamma_{n+1}^{(3)}/\Gamma_{n+1}$, $\Gamma_{n+1}^{(6)}/\Gamma_{n+1}$ which decay to zero faster than the 'standard' term

$\Gamma^{(2)}_{n+1}/\Gamma_{n+1}$. In other words, the effects of applying a second-order retraction will diminish asymptotically.

Nevertheless, the above results show that similar to the geodesic schemes, if we set the step sizes $\eta_k = \rho/\sqrt{k+k_0}$, then the SA scheme (2) using a first and/or second-order retraction finds an $\mathcal{O}(b_\infty + \log(n)/\sqrt{n})$-stationary point to (1) in $\mathcal{O}(n)$ iterations.

We consider now the Markovian setting and we adopt the same set of assumptions on the Markov chain as in **MA1**. Define the constants:

$$C^{\mathrm{Ret}}_{\hat{e}} = \{L_{\hat{e}}\bar{c} + L\hat{e}_\infty\}(e_\infty + b_\infty + 1) + 2L(b^2_\infty + e^2_\infty)\,, \quad C^{\mathrm{Ret}}(\eta_1) = \bar{c}\hat{e}_\infty(2+\eta_1)\,,$$
$$D^{\mathrm{Ret}} = \bar{c}\hat{e}_\infty(a_2+1) + L_{\hat{e}}\bar{c}(a_1(b_\infty + e_\infty)+1) + L\hat{e}_\infty, \quad \widetilde{e}^{(i)} = h^i_\infty + b^i_\infty + e^i_\infty,\ i \in \mathbb{N}^*\,.$$

Theorem 6. *Assume **A1-A2-A3-A4**, **R1-R2-R3**, **MA1** hold. Assume that $(\eta_k)_{k\in\mathbb{N}^*}$ is a sequence of stepsizes and $a_1, a_2 \geq 0$ satisfying*

$$\sup_{k\in\mathbb{N}^*}\{\eta_{k+1}/\eta_k\} \leq 1\,, \qquad \sup_{k\in\mathbb{N}^*}\{\eta_k/\eta_{k+1}\} \leq a_1\,, \quad \sup_{k\in\mathbb{N}^*}\{|\eta_k - \eta_{k+1}|/\eta_k^2\} \leq a_2\,, \tag{14}$$

Consider $(\theta_n)_{n\in\mathbb{N}}$ defined by (2). For any $n \in \mathbb{N}^*$, let $I_n \in \{0,\ldots,n\}$ be distributed with (4).

*(a) Assume **R4**. If in addition $\sup_{k\in\mathbb{N}^*}\eta_k \leq (4L + 2D^{\mathrm{Ret}} + 6V_\infty\mathscr{L}^{(1)}_\infty)^{-1}\underline{c}$, then*

$$\mathbb{E}\left[\|h(\theta_{I_n})\|^2_{\theta_{I_n}}\right] \leq 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + C^{\mathrm{Ret}}(\eta_1) + c_0\Gamma^{(2)}_{n+1} + c_1\Gamma^{(3)}_{n+1} + c_2\Gamma^{(4)}_{n+1}\right\} + 2V_\infty b_\infty/\underline{c}\,, \tag{15}$$

where $c_0 = C^{\mathrm{Ret}}_{\hat{e}} + (3\mathscr{L}^{(1)}_\infty V_\infty)(b^2_\infty + e^2_\infty)$, $c_1 = 3\hat{e}_\infty\mathscr{L}^{(1)}_\infty(L_{\hat{e}}h_\infty + L)\widetilde{e}^{(2)}$, $c_2 = 54L(\mathscr{L}^{(1)}_\infty)^2\widetilde{e}^{(4)}$.

*(b) Assume **R5**. If in addition $\sup_{k\in\mathbb{N}^*}\eta_k \leq (4L + 2D^{\mathrm{Ret}})^{-1}\underline{c}$, then*

$$\mathbb{E}\left[\|h(\theta_{I_n})\|^2_{\theta_{I_n}}\right] \leq 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + C^{\mathrm{Ret}}(\eta_1) + c_0\Gamma^{(2)}_{n+1} + c_1\Gamma^{(3)}_{n+1} + c_2\Gamma^{(6)}_{n+1}\right\} + 2V_\infty b_\infty/\underline{c}\,, \tag{16}$$

where $c_0 = C^{\mathrm{Ret}}_{\hat{e}}$, $c_1 = 9\mathscr{L}^{(2)}_\infty(\hat{e}_\infty(L_{\hat{e}}h_\infty + L) + V_\infty)\widetilde{e}^{(3)}$, $c_2 = 486L(\mathscr{L}^{(2)}_\infty)^2\widetilde{e}^{(6)}$.

The proof is postponed to Appendix D.3.

As expected, the convergence of the retraction scheme with Markov noise demonstrate an analogous behavior as in the martingale noise setting. Particularly, using a step size of $\eta_k = \rho/\sqrt{k+k_0}$, the retraction scheme finds an $\mathcal{O}(b_\infty + \log(n)/\sqrt{n})$-stationary point to (1) in $\mathcal{O}(n)$ iterations. Furthermore, we observe that the step size conditions and constants are improved when a second-order retraction is used in lieu of a first-order one.

Examples of Retraction Maps We present retraction maps on two common Riemannian manifolds. Importantly, we show that they satisfy the assumptions stated in Theorems 5 and 6.

Projective Retraction on the Sphere – Consider the Euclidean unit sphere manifold $\Theta = \mathbb{S}^d = \{x \in \mathbb{R}^{d+1} : \|x\| = 1\}$, where $\|\cdot\|$ stands for the standard Euclidean norm on $\mathbb{R}^{d+1}$. By [2, Example 3.5.1] for any $\theta \in \mathbb{S}^d$, $\mathrm{T}_\theta\mathbb{S}^d = \{u \in \mathbb{R}^{d+1} : u^\top\theta = 0\}$. The Riemannian metric g is the canonical metric on the sphere, defined as the restriction of Euclidean scalar product from $\mathbb{R}^{d+1}$ to the tangent space $\mathrm{T}_\theta\mathbb{S}^d$. The corresponding Riemannian exponential is given by:

$$\mathrm{Exp}_\theta(u) = \cos(\|u\|)\,\theta + \sin(\|u\|)(u/\|u\|)\ . \tag{17}$$

The following result holds.

Proposition 7. *The projective retraction* Ret_θ *defined for any* $(\theta, u) \in \mathrm{T}\mathbb{S}^d$ *by*

$$\mathrm{Ret}_\theta(u) = (\theta + u)/\|\theta + u\| \tag{18}$$

satisfies ***R****1*–***R****5 and for any* $\theta \in \Theta$, Φ_θ *defined by* (11) *has a bounded first derivative.*

The proof is postponed to Appendix D.4.

The retraction (18) is *both* a first-order and second-order retraction, and Φ_θ also has a bounded first-order derivative. Consequently, Theorem 5, & 6 & 14 can be applied according to conditions on the noise properties. We remark that by comparing (17) with (18), the retraction map Ret_θ has a better numerical stability as it does not involve evaluating the trigonometric functions.

Projective Retraction on the Grassmannian – Consider $\Theta = \mathrm{Gr}_r(\mathbb{R}^d)$ as the real Grassmann manifold with its quotient space Riemannian metric [16, Section 2.3.2]. The manifold $\mathrm{Gr}_r(\mathbb{R}^d)$ is the set of r-dimensional subspaces of $\mathbb{R}^d$. Each $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$ can be represented by some $B \in \mathrm{St}_r(\mathbb{R}^d) = \{B \in \mathbb{R}^{d\times r} : B^\top B = \mathrm{I}_r\}$, where $\mathrm{St}_r(\mathbb{R}^d)$ is the Stiefel manifold [2, Section 3.3.2]. Each element $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$ is then denoted as $\theta = [B] = \mathrm{Span}(B)$. By [16, Section 2.5], for any $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$, with representative $B \in \mathrm{St}_r(\mathbb{R}^d)$, the tangent space $\mathrm{T}_\theta\mathrm{Gr}_r(\mathbb{R}^d)$ is given by

$$\mathrm{T}_\theta\mathrm{Gr}_r(\mathbb{R}^d) = \left\{u = B_\perp C \,:\, C \in \mathbb{R}^{(d-r)\times r}\right\}\ , \tag{19}$$

where $B_\perp \in \mathrm{St}_{d-r}(\mathbb{R}^d)$ is such that $[B_\perp]$ is the orthogonal complement of $\theta = [B]$. The exponential map on $\mathrm{Gr}_r(\mathbb{R}^d)$ corresponding to the metric $\mathrm{g}_\theta(u, v) = \mathrm{Tr}(u^\top v)$ for any $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$, $u, v \in \mathrm{T}_\theta\mathrm{Gr}_r(\mathbb{R}^d)$, is

$$\mathrm{Exp}_\theta(u) = \left[(B\,, B_\perp)\ \exp\begin{pmatrix} 0 & -C^\top \\ C & 0 \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ 0_{d-r\times r} \end{pmatrix}\right]\,. \tag{20}$$

The following results hold.

Proposition 8. *The projective retraction* Ret *defined for any* $(\theta, u) \in \mathrm{T}_\theta \mathrm{Gr}_r(\mathbb{R}^d)$*, by*

$$\mathrm{Ret}_\theta(u) \;=\; [B+u] \tag{21}$$

satisfies $\boldsymbol{R}\,1$*–*$\boldsymbol{R}\,5$*. On the other hand, if* $r>1$*,* Φ_θ *does not have bounded first derivative.*

The proof is postponed to Appendix D.5.

Similar to the projective retraction on spheres, the retraction (21) is *both* a first-order and second-order retraction, and the results Theorem 5 & 6 follow. We remark that by comparing (20) with (21), the retraction map is simpler to compute as it involves only a simple matrix addition.

5 Applications

In this section, we illustrate our convergence analysis results on three application examples of Riemannian SA scheme. They are subspace tracking method, stochastic semidefinite programming via low rank reparameterization, and robust barycenter problem.

5.1 Principal Component Analysis by Subspace Tracking

We consider a principal component analysis (PCA) problem in which we look for the r principal eigenvectors of a $d \times d$ covariance matrix $\mathbf{A}$, for which we have access to noisy data $(X_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^d$. Note that this problem has been tackled in several works, see e.g. [30, 9, 43] and the references therein. PCA corresponds to the following minimization problem on the parameter space $\Theta = \mathrm{Gr}_r(\mathbb{R}^d)$, introduced above,

$$\min_{\theta=[B]\in\mathrm{Gr}_r(\mathbb{R}^d)} \left\{ f(\theta) = -\operatorname{Tr}(B^\top \mathbf{A} B) \middle/ 2 \right\} \;.$$

The mean field corresponds to $h = \operatorname{grad} f : \mathrm{Gr}_r(\mathbb{R}^d) \to \mathrm{TGr}_r(\mathbb{R}^d)$ and is given by, for any $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$,

$$\operatorname{grad} f(\theta) = (\mathrm{I}_d - BB^\top)\mathbf{A}B = -B_\perp B_\perp^\top \mathbf{A} B \;.$$

By writing $C = -B_\perp^\top \mathbf{A} B \in \mathbb{R}^{(d-r)\times r}$, note that we do have for any $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$, $\operatorname{grad} f(\theta) = B_\perp C \in \mathrm{T}_\theta\mathrm{Gr}_r(\mathbb{R}^d)$. Since f is infinitely differentiable and $\mathrm{Gr}_r(\mathbb{R}^d)$ is compact, $\mathbf{A}\,2$ is satisfied with $V = -f$. Therefore, using Proposition 8, if we define for any $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$,

$$H_\theta(x) = B_\perp(-B_\perp^\top x x^\top B) \;,$$

we can apply the results of Section 3 & 4 corresponding to either geodesic or retraction schemes for the sequence $(\theta_n)_{n\in\mathbb{N}}$ defined by (2), depending on the conditions of the data $(X_n)_{n\in\mathbb{N}}$.

Compared to the recent line of Riemannian based analysis of PCA, [9, 43], our analysis is more flexible. Particularly, it holds for a general retraction scheme, in lieu of the computationally expensive exponential map.

5.2 Semidefinite Programming

Consider the semidefinite programming problem,

$$\min_{C\in\mathbb{R}^{d\times d}} \ \mathrm{Tr}(AC)\ ,\text{s.t.}\ \ C = C^\top,\ \mathrm{diag}(C) = \mathbf{1_d},\ \mathbf{C} \succeq \mathbf{0}\ , \tag{22}$$

where $A \in \mathbb{R}^{d\times d}$ is a symmetric (possibly non positive semidefinite) matrix, $\mathbf{1_d} = (\mathbf{1},\ldots,\mathbf{1}) \in \mathbb{R}^{\mathbf{d}}$, and $\mathrm{diag}(B) = (B_{1,1},\ldots,B_{d,d})$ for any matrix $B = (B_{i,j})_{i,j\in\{1,\ldots,d\}}$. Problem (22) is a convex relaxation of several NP-hard problems such as MAXCUT [20] and community detection [1]. In these applications, A corresponds to the (weighted) adjacency or Laplacian matrix of a graph with d nodes. When d is large, solving (22) entails a high complexity since the SDP comprises of $\mathcal{O}(d^2)$ unknowns.

As shown in [31, 11], for some instances, there exists a low-rank optimal solution $C^\star$ to (22) with $\mathrm{rank}(C^\star) \leq p$ for some $p \ll d$. This motivates the re-parameterization, $C = \theta\theta^\top$, $\theta \in \mathbb{R}^{d\times p}$ and therefore to consider the non-convex quadratic program,

$$\min_{\theta\in\mathbb{R}^{d\times p}} \left\{ f(\theta) = \mathrm{Tr}(A\theta\theta^\top) \right\} \ \ \text{s.t.}\ \ \mathrm{diag}(\theta\theta^\top) = \mathbf{1_d}\ . \tag{23}$$

Due to the constraints of (23), this problem is therefore is equivalent to consider the minimization of f on the Riemannian manifold $\Theta = \prod_{i=1}^d \mathbb{S}^{p-1}$ endowed with the product metric induced by the canonical one on $\mathbb{S}^{p-1}$, see [28, p. 20]. It is easy to show that gradient $\mathrm{grad}\, f(\theta) \in \mathrm{T}_\theta\Theta$ is given by

$$\left[\mathrm{grad}\, f(\theta)\right]_{i,:} = c_i^\top \theta(\mathrm{I}_p - [\theta]_{i,:}([\theta]_{i,:})^\top) \qquad \text{for } i = 1, ..., d\ , \tag{24}$$

where $[\theta]_{i,:}$ denotes the ith row vector of θ. In addition, similarly to the sphere we define the projective retraction Ret on $\mathrm{T}\Theta$ for any $(\theta, u) \in \mathrm{T}\mathbb{S}^d$ by

$$[\mathrm{Ret}_\theta(u)]_{i,:} = (\theta_{i,:} + u_{i,:})/ \left\|\theta_{i,:} + u_{i,:}\right\|\ , \qquad \text{for } i = 1, ..., d\ , \tag{25}$$

Following the same lines as the proof of Proposition 7, Ret defined by (25) satisfies **R**1–**R**5 and for any $\theta \in \Theta$, Φ_θ defined by (11) has a bounded first derivative. Note that [11] analyze a deterministic version of the retraction scheme (2) that we consider to tackle (23). However, an exact gradient $\mathrm{grad}\, f$ is required in their algorithm.

Evaluating the exact gradient (24) is computationally challenging when d is large and A is a dense matrix. As a remedy, a natural idea is to select $(A_n)_{n\in\mathbb{N}^*}$ as $A_n = A \odot X_n/\delta$, where $X_n \in \{0,1\}^{d\times d}$ is a binary matrix with i.i.d. elements as $\mathbb{P}\left((X_n)_{1,1} = 1\right) = \delta$, for $\delta > 0$, and $\odot$ denotes element-wise product.

Since f is infinitely differentiable and $\mathrm{Gr}_r(\mathbb{R}^d)$ is compact, **A**2 is satisfied with $V = -f$. Therefore, if we define for any $\theta \in \mathrm{Gr}_r(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$,

$$\left[H_\theta(x)\right]_{i,:} = [x]_{i,:}\, \theta\, (\mathrm{I}_p - [\theta]_{i,:}([\theta]_{i,:})^\top)\ , \qquad \text{for } i = 1, ..., d\ ,$$

we can apply the results of Section 3 & 4 corresponding to either geodesic or retraction schemes for the sequence $(\theta_n)_{n\in\mathbb{N}}$ defined by (2), depending on the conditions we assume on the data $(A_n)_{n\in\mathbb{N}^*}$.

5.3 Robust Barycenter in a Hadamard Manifold

Let Θ be a Hadamard manifold – a simply-connected, complete Riemannian manifold of non-positive sectional curvature [28]. We assume that the sectional curvature of Θ is bounded below, $-\kappa^2 \leq \sec\Theta \leq 0$. A common example of this situation is $\Theta = \mathrm{S}_d^{++}$, the space of $d \times d$ symmetric positive-definite matrices, equipped with its affine-invariant metric [32] for which it holds $\kappa^2 = (1/8)\, d(d-1)(d+2)$ [4].

Consider a set of data points $(X_n)_{n\in\mathbb{N}^*}$ lying on the Riemannian manifold Θ, i.e., $X_n \in \Theta$, that are drawn from a distribution π. A fundamental machine learning problem is to compute some kind of central value of π, defined as an optimal solution to

$$\min_{\theta\in\Theta} \left\{ V(\theta) \,=\, \textstyle\int_\Theta \tilde{\rho}(\theta, x)\, \pi(\mathrm{d}x) \right\} \,, \tag{26}$$

where $\tilde{\rho}(\theta, x)$ is some Riemannian dissimilarity measure. For instance, the Riemannian barycenter, also called the Fréchet mean, is obtained by taking $\tilde{\rho}(\theta, x) = \rho^2_\Theta(\theta, x)$ where $\rho_\Theta : \Theta \times \Theta \to \mathbb{R}_+$ is the Riemannian distance of Θ [21, 29].

The Riemannian barycentre is known to be sensitive to outliers, which motivated the idea of considering the Riemannian median, obtained by taking $\tilde{\rho}(\theta, x) = \rho_\Theta(\theta, x)$ [5]. We consider a *robust barycenter* by using a Huber-like dissimilarity measure, $\tilde{\rho} : \Theta \times \Theta \to \mathbb{R}_+$,

$$\tilde{\rho}(\theta, \tau) \,=\, \delta^2 \left[1 + \{\rho_\Theta(\theta, \tau)/\delta\}^2 \right]^{1/2} - \,\delta^2 \,, \tag{27}$$

where $\delta > 0$ is a cut-off constant. Observe that $\tilde{\rho}(\theta, \tau)$ behaves like $(1/2)\, \rho^2_\Theta(\theta, \tau)$ when $\rho_\Theta(\theta, \tau)$ is small compared to δ, and like $\delta\, \rho_\Theta(\theta, \tau)$ when $\rho_\Theta(\theta, \tau)$ is large compared to δ. Let π be a probability distribution on Θ. Using (27) in the optimization problem (26) yields a *robust barycenter* problem, and the robust barycenter is a global minimizer of (26).

In the simplest setting where $(X_n)_{n\in\mathbb{N}^*}$ are i.i.d. from π, tackling the problem (26) can be done by consider the following geodesic SA scheme:

$$\theta_{n+1} \,=\, \mathrm{Exp}_{\theta_n}\left(\eta_{n+1} \, \frac{\mathrm{Exp}^{-1}_{\theta_n}(X_{n+1})}{\left[1+\{\rho_\Theta(\theta_n\,, X_{n+1})/\delta\}^2\,\right]^{1/2}} \right) \,. \tag{28}$$

The above is essentially the "recursive barycenter" scheme proposed by [36, 6], except that a move in the direction of a new observation X_{n+1} is attenuated when this new observation lies too far from the current estimate θ_n . This means that less confidence is assigned to extreme observations. We show in Appendix E that results from Section 3 can be applied and furthermore (28) finds a unique and global minimizer to (26). Note that our analysis can be easily extended to the case when $(X_n)_{n\in\mathbb{N}^*}$ is a Markov chain with stationary distribution of π and satisfying **MA**1.

Conclusions In this paper, we have provided a comprehensive study for the convergence analysis of Riemmanian SA schemes. For applications to possibly non-convex optimization problems, our results generalize a number of recent works in the sense that the SA scheme is potentially biased and we allow the general retraction to be used in lieu of the complex geodesic operation. We also demonstrate our results on three examples motivated by machine learning applications.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] P. Absil, R. Mahony, and R. Sepulchre. *Optimisation algorthms on matrix manifolds*. Princeton University Press, 2008.

[3] P. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

[4] A. Andai. *Information Geometry in quantum mechanics*. PhD thesis, Budapest University of Technology and Economics, 2003.

[5] M. Arnaudon, F. Barbaresco, and L. Yang. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604, 2013.

[6] M. Arnaudon, C. Dombry, A. Phan, and L. Yang. Stochastic algorithms for computing means of probability measures. *Stochastic Processes and their Applications*, 122(4):1437 – 1455, 2012.

[7] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350, Nov. 2001.

[8] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

[9] S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[10] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2018.

[11] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

[12] O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):593–613, 2009.

[13] C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In *NeurIPS*, 2019.

[14] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains.* Springer, 2018.

[15] M. Duflo. *Random Iterative Models.* Springer-Verlag, Berlin, Heidelberg, 1st edition, 1997.

[16] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[17] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2011.

[18] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian geometry (Third Edition).* Springer-Verlag, 2004.

[19] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[20] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[21] A. Goh and R. Vidal. Clustering and dimensionality reduction on riemannian manifolds. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.

[22] R. Hosseini and S. Sra. An alternative to em for gaussian mixture models: Batch and stochastic riemannian optimization. *Mathematical Programming*, pages 1–37, 2019.

[23] T. Jaakkola, M. I. Jordan, and S. P. Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 703–710, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[24] J. Jost. *Riemannian Geometry and Geometric Analysis.* Springer Universitat texts. Springer, 2005.

[25] B. Karimi, B. Miasojedow, E. Moulines, and H.-T. Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 1944–1974, 2019.

[26] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[27] J. LaSalle. Some extensions of liapunov's second method. *IRE Transactions on circuit theory*, 7(4):520–527, 1960.

[28] J. Lee. *Introduction to Riemannian Manifolds*. Springer International Publishing, 2019.

[29] E. Mathieu, C. Le Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh. Continuous hierarchical representations with poincaré variational auto-encoders. In *Advances in neural information processing systems*, pages 12544–12555, 2019.

[30] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927 – 935, 1992.

[31] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

[32] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[33] H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407, 1951.

[34] T. Sakai. On the cut loci of compact symmetric spaces. *Hokkaido Mathematical Journal*, 6(3):136–161, 1976.

[35] S. M. Shah. Stochastic approximation on riemannian manifolds. *Applied Mathematics & Optimization*, May 2019.

[36] K. T. Sturm. Probability measures on metric spaces of nonpositive curvature. *Contemporary Mathematics*, 338, 01 2003.

[37] T. Sun, Y. Sun, and W. Yin. On markov chain gradient descent. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9896–9905. Curran Associates, Inc., 2018.

[38] Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on riemannian manifolds. In *NeurIPS*, 2019.

[39] J. Townsend. Differentiating the singular value decomposition. Technical report, University College London, 2016.

[40] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on riemannian manifolds. In *Conference On Learning Theory, COLT*, pages 650–687, 2018.

[41] C. Udriste. *Convex functions and optimisation methods on Riemannian manifolds*. Springer Science + Business Media, 1994.

[42] H. Zhang, S. J. Reddi, and S. Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600, 2016.

[43] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory, COLT*, pages 1617–1638, 2016.

[44] J. Zhang, H. Zhang, and S. Sra. R-spider: A fast riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*, 2018.

A Preliminaries on Riemannian Geometry

This section reviews some basic concepts about Riemannian geometry. These concepts are essential to develop our main results on convergence of Riemannian SA.

A.1 Metric Tensor and Distance

A smooth manifold (at least C^2) Θ is equipped with a smooth metric tensor field $g \in T^2\mathrm{T}^*\Theta$, see [28, Proposition 2.4]. To each $\theta \in \Theta$, this associates a scalar product g_θ on the tangent space $\mathrm{T}_\theta\Theta$. When there is no confusion, we denote that [28, Chapter 2, pages 11-12],

$$g_\theta(u, v) = \langle u, v \rangle_\theta \ , \quad u, v \in \mathrm{T}_\theta\Theta$$

and the corresponding norm on $\mathrm{T}_\theta\Theta$ is called the Riemannian norm, $\|u\|_\theta = \langle u, u \rangle_\theta^{1/2}$.

With the metric tensor, it is possible to define the notion of length of a curve. If $c : I \to \Theta$ is a differentiable curve, defined on some interval $I \subset \mathbb{R}$, with velocity $\dot{c}$, then its length is [28, page 34]

$$\ell(c) \ = \int_I \|\dot{c}(t)\|_{c(t)} \ \mathrm{d}t \ . \tag{29}$$

The length $\ell(c)$ is invariant by reparametrization: $\ell(c \circ \phi) = \ell(c)$ for any diffeomorphism $\phi : J \to I$, from an interval J onto I. Thus, without loss of generality, we consider only curves that are restricted to $c : [0, 1] \to \Theta$.

This can be used to turn Θ into a metric space. Indeed, if Θ is connected, the following is a well-defined distance function; satisfying the axioms of a metric space [28, Theorem 2.55],

$$\rho_\Theta(\theta, \theta') \ = \ \inf \left\{ \ell(c) \,|\, c : [0, 1] \to \Theta \,;\, c(0) = \theta \,, c(1) = \theta' \right\} \ , \quad \theta, \theta' \in \Theta \ . \tag{30}$$

This is called the Riemannian distance induced by the metric tensor g.

The infimum in (30) is always attained for any $\theta, \theta' \in \Theta$, provided that the distance $\rho_\Theta(\cdot, \cdot)$ turns Θ into a complete metric space. This is a corollary of the Hopf-Rinow theorem, a fundamental theorem in Riemannian geometry [28, Corollary 6.21].

A.2 Levi-Civita Connection

In the Riemannian mainfold, the curve γ which attains the infimum in (30) is called a geodesic. Intuitively, a geodesic is a C^2 curve which has zero acceleration. This intuition can be formalized by introducing an affine connection ∇ [28, page 89], compatible with the metric tensor g, called the Levi-Civita connection, or just Riemannian connection (to be precise, $\nabla = \nabla^g$, depends on the choice of g).

To each vector $u \in \mathrm{T}_\theta\Theta$ and smooth vector field X on Θ, the connection ∇ associates a vector $\nabla_u X \in \mathrm{T}_\theta\Theta$. This vector is called the covariant derivative of X in the direction of u. This is bilinear in u and X, and satisfies the product rule

$$\nabla_u(fX) = (uf)\, X(\theta) + f(\theta)\, \nabla_u X \ , \tag{31}$$

for any differentiable function $f : \Theta \to \mathbb{R}$, where uf denotes the derivative of f along u. Moreover, one has the following,

$$u \,\langle X, Y\rangle \ = \ \langle \nabla_u X, Y\rangle_\theta + \langle X, \nabla_u Y\rangle_\theta \ , \tag{32}$$

$$\nabla_X Y - \nabla_Y X = [X, Y] \ , \tag{33}$$

for any differentiable vector fields X, Y on Θ, where $[X, Y]$ is the Lie bracket of the vector fields X and Y, itself a vector field. Here, (32) states that ∇ is compatible with the metric, and (33) states that ∇ is a connection with zero torsion.

The Levi-Civita connection [28, Theorem 5.10] is defined as the unique affine connection ∇ which satisfies (32) and (33). Note that the uniqueness of this connection can be guaranteed by the Koszul's theorem, also known as the fundamental theorem of Riemannian geometry.

A.3 Geodesic Equation

It can be proved using (31) that $\nabla_u X$ depends only on the values of X along a curve tangent to the vector u [28, Proposition 4.26]. This motivates the following definition. Consider $c : I \to \Theta$ as a smooth curve on Θ and X is an extendible vector field along c, this means that $X : I \to \mathrm{T}\Theta$ satisfies $X(t) \in \mathrm{T}_{c(t)}\Theta$ for any $t \in I$, see [28, pages 100-101]. The covariant derivative of X along c is defined by

$$\mathrm{D}_t X = \nabla_{\dot{c}} \tilde{X} \circ c \ , \tag{34}$$

where $\tilde{X}$ is a vector field on Θ satisfying, for any $t \in I$, $X(t) = \tilde{X} \circ c(t)$. The reader should not confuse the index t in D_t, which is just a notation, with an actual real number $t \in I$.

A *geodesic* is thus a smooth curve $\gamma : I \to \Theta$, whose velocity $\dot{\gamma}$ is parallel along γ. If D_t is the covariant derivative along γ, then γ satisfies the geodesic equation [28, page 103],

$$\mathrm{D}_t \dot{\gamma}(t) = 0 \ , \quad t \in I \ . \tag{35}$$

The left-hand side of this equation is precisely the acceleration of the curve γ.

The geodesic equation is a non-linear ordinary differential equation of second order. For given initial conditions $\gamma(0) = \theta$ and $\dot{\gamma}(0) = u$, it has a unique solution $\gamma : (-\varepsilon, \varepsilon) \to \Theta$, for some $\varepsilon > 0$ [28, Theorem 4.27]. If this solution can always be extended to a curve $\gamma : \mathbb{R} \to \Theta$, then Θ is called a complete Riemannian manifold. The Hopf-Rinow theorem states that this is equivalent to Θ being a complete metric space, with the distance function (30) [28, Theorem 6.19].

A.4 Parallel Transport and Parallel Frames

Recall D_t the covariant derivative, associated with the Levi-Civita connection, along a curve $c : [0, 1] \to \Theta$, given in (34). Then, a vector field X is said to be parallel along c if it satisfies the parallel transport equation

$$\mathrm{D}_t X(t) = 0 \; .$$

This is a first-order linear ordinary differential equation (ODE). Say $c(0) = \theta_0$, then, for a given initial condition $X(0) = u$, where $u \in \mathrm{T}_{\theta_0}\Theta$, it follows that $u \mapsto X(t)$ is a linear mapping from $\mathrm{T}_{\theta_0}\Theta$ to $\mathrm{T}_{c(t)}\Theta$ [28, Theorem 4.32]. This is denoted T^c_{0t}, and by uniqueness of the solution to the ODE, T^c_{t0} is its linear inverse [28, Equation (4.22)].

It is useful to derive an equivalent condition to (32), which holds for vector fields X and Y along c [28, Proposition 5.5].

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\langle X, Y \right\rangle_{c(t)} = \left\langle \mathrm{D}_t X, Y \right\rangle_{c(t)} + \left\langle X, \mathrm{D}_t Y \right\rangle_{c(t)} \; . \tag{36}$$

This equation yields that $t \mapsto \langle X, Y\rangle_{c(t)}$ is constant if X and Y are parallel vector fields along c. Thus T^c_{0t} preserves scalar products. In particular, if $(\mathbf{b}_i \,; i = 1, \ldots, d)$ is an orthonormal basis of $\mathrm{T}_{\theta_0}\Theta$, then the vector fields along c, defined by

$$e_i(t) = \mathrm{T}^c_{0t}\, \mathbf{b}_i \; ,$$

form an orthonormal basis of the tangent space $\mathrm{T}_{c(t)}\Theta$, for each $t \in I$. This is called a parallel orthonormal frame along c [28, Equation (4.23)]. By linearity of T^c_{0t}, if $u \in T_{\theta_0}\Theta$ is written $u = \sum_{i=1}^d u^i\, \mathbf{b}_i$, then

$$\mathrm{T}^c_{0t} u \; = \sum_{i=1}^{d} u^i\, e_i(t) \; . \tag{37}$$

In other words, parallel transport is obtained by simply propagating the cooordinates u^i of the vector u along a parallel orthonormal frame.

A.5 Riemannian Exponential Map and Cut Locus

From now on, let us assume that Θ is a complete Riemannian manifold. A curve that attains the infimum in (30) is called a length-minimizing curve. While this curve is not always unique, it is always a geodesic [28, Theorem 6.4]; in other words, it is a twice differentiable solution of the geodesic equation (35). On the other hand, it is very important to keep in mind that a geodesic is not always a length-minimizing curve.

To give a concrete example, consider the geodesics of a sphere with its usual round metric [28, Example 2.13] which are simply its great circles, i.e., intersections of the sphere with planes passing through the origin. Clearly, a portion of a great circle whose length is greater than π is not length-minimizing. Therefore, geodesics which start at some point θ on a sphere, are length-minimizing until they reach the opposite point $-\theta$. One says that the cut locus of the point θ on the sphere is the set $\{-\theta\}$.

Since Θ is complete, for $u \in \mathrm{T}_\theta\Theta$, there exists a unique geodesic $\gamma_u : \mathbb{R} \to \Theta$ with $\gamma_u(0) = \theta$ and $\dot{\gamma}_u(0) = u$. Then, [28, page 128] define the *exponential map* $\mathrm{Exp} : \mathrm{T}\Theta \to \Theta$ as:

$$\mathrm{Exp}_\theta(u) = \gamma_u(1) \ , \tag{38}$$

which is a smooth map. For $u \in \mathrm{T}_\theta\Theta$, with $\|u\|_\theta = 1$, let $c(u) > 0$ be the largest positive number t such that γ_u is length-minimizing when restricted to the interval $[0, t]$ [28, page 307]:

$$c(u) = \sup\{t \geq 0 \,:\, \rho_\Theta(\theta, \gamma_u(t)) = t\} \ . \tag{39}$$

The tangent cut locus of θ is then defined by

$$\mathrm{TCut}(\theta) \ = \ \{c(u)u \,:\, u \in \mathrm{T}_\theta\Theta\, ;\, \|u\|_\theta = 1\} \ .$$

Finally, the cut locus of θ is given by [28, page 308]

$$\mathrm{Cut}(\theta) \ = \ \mathrm{Exp}_\theta\{\mathrm{TCut}(\theta)\} \ .$$

For example, if Θ is a unit sphere, then $\mathrm{TCut}(\theta)$ is the set of tangent vectors $u \in \mathrm{T}_\theta\Theta$ such that $\|u\|_\theta = \pi$. On the other hand, $\mathrm{Cut}(\theta) = \{-\theta\}$, since $\|u\|_\theta = \pi$ implies $\mathrm{Exp}_\theta(u) = -\theta$.

A.6 Injectivity Domain

The cut locus of a point in a complete Riemannian manifold gives valuable information regarding the topology of the manifold. Indeed, if Θ is a complete Riemannian manifold and θ is any point in Θ, then Θ can be decomposed into the disjoint union of two sets

$$\Theta = \mathrm{D}(\theta) \cup \mathrm{Cut}(\theta) \ , \tag{40}$$

where $\mathrm{D}(\theta)$ is the injectivity domain,

$$\mathrm{D}(\theta) \ = \ \mathrm{Exp}_\theta\{\mathrm{TD}(\theta)\} \ , \quad \text{where } \mathrm{TD}(\theta) = \{tu \,:\, u \in \mathrm{T}_\theta\Theta\, ;\, \|u\|_\theta = 1 \text{ and } 0 \leq t < c(u)\} \ ,$$

where $c(u)$ is given by (92). We observe:

Proposition 9. *[28, Theorem 10.34] The Riemannian exponential map* Exp_θ *is a diffeomorphism of* $\mathrm{TD}(\theta)$ *onto* $\mathrm{D}(\theta)$*. Therefore, the inverse of the Riemannian exponential* Exp_θ^{-1} *is well-defined, and a diffeomorphism, on* $\mathrm{D}(\theta) = \Theta - \mathrm{Cut}(\theta)$*.*

In fact, $\mathrm{TD}(\theta)$ is an open, star-shaped subset of the tangent space $\mathrm{T}_\theta\Theta$, so it has the topology of an open ball. Thus, (40) states that the topology of Θ is completely determined by $\mathrm{Cut}(\theta)$. This theorem also ensures that $\mathrm{Cut}(\theta)$ is a closed set of measure zero.

A.7 Isometries and Homogeneous Spaces

An isometry g on the Riemannian manifold Θ is a diffeomorphism $g : \Theta \to \Theta$ which preserves the Riemannian metric. To express this, let $g \cdot \theta = g(\theta)$, and $g \cdot u = \mathrm{D}g_\theta(u)$ for each $\theta \in \Theta$ and $u \in \mathrm{T}_\theta\Theta$. Here, $g(\theta) \in \Theta$ is simply the image of θ under the map g, and $\mathrm{D}g_\theta$ denotes the derivative of g at θ, so $Dg_\theta : \mathrm{T}_\theta\Theta \to \mathrm{T}_{g\cdot\theta}\Theta$. We say that g is an isometry if [28, page 12]

$$\langle g \cdot u, g \cdot v\rangle_{g\cdot\theta} = \langle u, v\rangle_\theta \qquad u\,,\, v \in \mathrm{T}_\theta\Theta\;.$$

In other words, the linear map $u \mapsto g \cdot u$ from $\mathrm{T}_\theta\Theta$ to $\mathrm{T}_{g\cdot\theta}\Theta$ preserves scalar products. In particular, it also preserves norms, so $\|g \cdot u\|_{g\cdot\theta} = \|u\|_\theta$.

Isometries also preserve objects derived from the Riemannian metric such as distance, geodesics, among others. In particular, if $\gamma : I \to \Theta$ is a geodesic, and $g : \Theta \to \Theta$ is an isometry, then $\gamma' = g \circ \gamma : I \to \Theta$ is also a geodesic [28, Corollary 5.14]. Now, if $\gamma(0) = \theta$ and $\dot{\gamma}(0) = u$, then $\gamma'(0) = g \cdot \theta$ and $\dot{\gamma}'(0) = g \cdot u$. From the definition of the Riemannian exponential (38), it is seen that

$$g \cdot \mathrm{Exp}_\theta(u) = \mathrm{Exp}_{g\cdot\theta}(g \cdot u)\;. \tag{41}$$

The set G of all isometries of a Riemannian manifold Θ forms a group under composition. A deep theorem, called Myers-Steenrod theorem, states that G can always be given the structure of a Lie group, such that for each $\theta \in \Theta$, the group action $g \mapsto g \cdot \theta$ is a differentiable map [18, page 66].

One calls Θ a Riemannian homogeneous space if its group of isometries G acts transitively. Transitive action means that for any $\theta, \theta' \in \Theta$ there exists $g \in \mathsf{G}$ such that $g \cdot \theta = \theta'$. When Θ is a Riemannian homogeneous space, knowing the metric of Θ at just one point, $o \in \Theta$, is enough to know this metric anywhere [18, page 67].

A.8 Riemannian Gradient, Hessian, and Taylor Formula

The metric tensor $\langle\cdot,\cdot\rangle_{\cdot}$ and Levi-Civita connection ∇, on the Riemannian manifold Θ, can be used to generalize classical objects from analysis, like the gradient and Hessian of a C^2 function $V : \Theta \to \mathbb{R}$, as we introduce next.

The Riemannian gradient of V is a vector field $\mathrm{grad}\,V$ on Θ, uniquely defined by the property [28, Equation 2.14]

$$\langle \mathrm{grad}\,V, u\rangle_\theta \;=\; \mathrm{D}V(\theta)(u)\;, \quad u \in T_\theta\Theta\;, \tag{42}$$

where $\mathrm{D}V(\theta) : \mathrm{T}_\theta\Theta \to \mathbb{R}$ is the differential of the function V at θ. As V is a real-valued function, it is useful to know that differentials, directional derivatives and covariant derivatives coincide $\mathrm{D}V(\theta)(u) = u\,V(\theta) = \nabla_u V(\theta)$. This definition makes it clear that the Riemannian gradient $\mathrm{grad}\,V$ depends on the choice of metric on the manifold Θ, and does not arise from the manifold structure of Θ, in itself.

The Riemannian Hessian of V, denoted $\operatorname{Hess} V$ is defined using the Levi-Civita connection. Precisely, it is the covariant derivative of the gradient $\operatorname{grad} V$. For $\theta \in \Theta$, this gives the Hessian $\operatorname{Hess} V(\theta) : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$,

$$\operatorname{Hess} V(\theta)\, u \;=\; \nabla_u \operatorname{grad} V(\theta)\;. \tag{43}$$

The Riemannian definition of the Hessian coincides with the covariant Hessian obtained from the Levi-Civita connection [28, Example 4.22]. The covariant characterization gives, for any vector fields X, Y on Θ,

$$\langle \operatorname{Hess} V X, Y\rangle = \nabla_X \nabla_Y V - \nabla_{\nabla_X Y} V = X(Y\,V) - (\nabla_X Y)V\;,$$

where the last equality comes from the remark regarding directional derivatives. One can see that (33) yields that the linear operator $\operatorname{Hess} V(\theta) : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$ is self-adjoint with respect to the Riemannian scalar product $\langle\cdot,\cdot\rangle_\theta$. Therefore, and not without a slight abuse of notation, we will also call Hess the resulting symmetric bilinear form

$$\operatorname{Hess} V(\theta)(u,v) = \langle \operatorname{Hess} V(\theta)u, v\rangle_\theta\;, \quad u, v \in \mathrm{T}_\theta\Theta.$$

Using the gradient (42) and the Hessian (43), one can derive the following Taylor formula for the function V. If $\gamma : [0,1] \to \Theta$ is a geodesic such that $\gamma(0) = \theta_0$ and $\gamma(1) = \theta_1$; which we simply call a geodesic between θ_0 and θ_1; then we have

$$V(\theta_1) - V(\theta_0) = \langle \operatorname{grad} V(\theta_0), \dot\gamma(0)\rangle_{\theta_0} + \operatorname{Hess} V(\gamma(t_*))\,(\dot\gamma,\dot\gamma)/2\;,$$

for some $t_* \in (0,1)$.

A.9 Bounded Hessian Implies Lipschitz Gradient

We can now state a result that can be very useful when the Riemannian Hessian is bounded.

Lemma 10. *If V has a continuous Riemannian Hessian* $\operatorname{Hess} V(\theta) : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$*, with operator norm upper bounded uniformly for $\theta \in \Theta$, say by N; then the Riemannian gradient* $\operatorname{grad} V$ *satisfies the Lipschitz property (3) with Lipschitz constant $L = N$.*

Proof. Let $(e_i; i = 1, \ldots, d)$ be a parallel orthonormal frame along γ. Define $\nabla V^i : [0,1] \to \mathrm{T}_\theta\Theta$ by

$$\nabla V^i(t) = \langle \operatorname{grad} V(\gamma(t)), e_i(t)\rangle_{\gamma(t)}\;, \quad t \in [0,1]\;. \tag{44}$$

Also, note from (37), applied to $c = \gamma$ and $u = \operatorname{grad} V(\theta_0)$, that

$$\mathrm{T}^\gamma_{01} \operatorname{grad} V(\theta_0) \;=\; \sum_{i=1}^{d} \nabla V^i(0)\, e_i(1)\;.$$

Then, since $(e_i(1); i=1,\dots,d)$ is an orthonormal basis of $\mathrm{T}_{\theta_1}\Theta$,

$$\|\mathrm{grad}\, V(\theta_1) - \mathrm{T}^{\gamma}_{01}\mathrm{grad}\, V(\theta_0)\|^2_{\theta_1} \;=\; \sum_{i=1}^{d} (\nabla V^i(1) - \nabla V^i(0))^2 \tag{45}$$

But, by applying (36) to $c=\gamma$ with $X = \mathrm{grad}\, V\circ\gamma$ and $Y=e_i$, it follows from (34), (43) and (44),

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla V^i(t) \;=\; \langle \mathrm{Hess}\, V(\gamma(t))\dot{\gamma}(t), e_i(t)\rangle_{\gamma(t)} + \langle \mathrm{grad}\, V(\gamma(t)), \mathrm{D}_t e_i(t)\rangle_{\gamma(t)}\ . \tag{46}$$

Then, since each $e_i(t)$ is parallel along γ, $\mathrm{D}_t e_i(t) = 0$. Plugging (46) into (45) yields, by the mean-value theorem, followed by Jensen's inequality,

$$\begin{aligned}
\|\mathrm{grad}\, V(\theta_1) - \mathrm{T}^{\gamma}_{01}\mathrm{grad}\, V(\theta_0)\|_{\theta_1} &= \left(\sum_{i=1}^{d}\left[\int_0^1 \langle \mathrm{Hess}\, V(\gamma(t))\dot{\gamma}(t), e_i(t)\rangle_{\gamma(t)}\,\mathrm{d}t\right]^2\right)^{1/2}\ , \\
&\le \int_0^1 \left(\sum_{i=1}^{d}\left[\langle \mathrm{Hess}\, V(\gamma(t))\dot{\gamma}(t), e_i(t)\rangle_{\gamma(t)}\right]^2\right)^{1/2}\mathrm{d}t\ , \\
&= \int_0^1 \|\mathrm{Hess}\, V(\gamma(t))\dot{\gamma}(t)\|_{\gamma(t)}\,\mathrm{d}t\ .
\end{aligned}$$

The last equality comes from the fact that $(e_i(t)\,;\, i=1,\dots,d)$ is an orthonormal basis of $\mathrm{T}_{\gamma(t)}\Theta$. Finally, using the definition of the operator norm $\|\cdot\|_{op,\gamma(t)}$ on $\mathrm{T}_{\gamma(t)}\Theta$, the bound in the assumption and (29),

$$\|\mathrm{grad}\, V(\theta_1) - \mathrm{T}^{\gamma}_{01}\mathrm{grad}\, V(\theta_0)\|_{\theta_1} \le \int_0^1 \|\mathrm{Hess}\, V(\gamma(t))\|_{op,\gamma(t)}\,\|\dot{\gamma}(t)\|_{\gamma(t)}\,\mathrm{d}t \le N\ell(\gamma)\ .$$

This concludes the proof. □

B Proof of Lemma 1

Using $\gamma(0)=\theta_0$, $\gamma(1)=\theta_1$, by a Taylor expansion and the definition of the Riemaniann gradient, we have

$$V(\theta_1) - V(\theta_0) = \int_0^1 \langle \mathrm{grad}\, V(\gamma(t)), \dot{\gamma}(t)\rangle_{\gamma(t)}\,\mathrm{d}t = \int_0^1 \langle \mathrm{grad}\, V(\gamma(t)), \mathrm{T}^{\gamma}_{0t}\dot{\gamma}(0)\rangle_{\gamma(t)}\,\mathrm{d}t\ ,$$

where we have used for the last equality the uniqueness of the parallel transport [28, Theorem 4.32] and that γ is a geodesic. Therefore, we obtain, using that the parallel transport is a linear isometry [28, Proposition 5.5], that

$$
\begin{aligned}
&\left|V(\theta_1) - V(\theta_0) - \langle \operatorname{grad} V(\theta_0), \dot{\gamma}(0)\rangle_{\gamma(0)}\right| \\
&\quad \leq \int_0^1 \left|\langle \operatorname{grad} V(\gamma(t)), \mathrm{T}_{0t}^{\gamma}\dot{\gamma}(0)\rangle_{\gamma(t)} - \langle \operatorname{grad} V(\gamma(0)), \dot{\gamma}(0)\rangle_{\gamma(0)}\right| \mathrm{d}t \;, \\
&\quad = \int_0^1 \left|\langle \operatorname{grad} V(\gamma(t)) - \mathrm{T}_{0t}^{\gamma}\operatorname{grad} V(\gamma(0)), \mathrm{T}_{0t}^{\gamma}\dot{\gamma}(0)\rangle_{\gamma(t)}\right| \mathrm{d}t \;, \\
&\quad \leq L\int_0^1 \ell\left(\gamma_{\,|[0,t]}\right) \left\|\mathrm{T}_{0t}^{\gamma}\dot{\gamma}(0)\right\|_{\gamma(t)} \mathrm{d}t \leq L\ell(\gamma)\int_0^1 \ell(\gamma)t\,\mathrm{d}t \;,
\end{aligned}
$$

where we have used that since γ is a geodesic $\mathrm{T}_{0t}^{\gamma}\dot{\gamma}(0) = \dot{\gamma}(t)$ and by [28, Corollary 5.6], $\|\dot{\gamma}(t)\|_{\gamma(t)} = \ell(\gamma)$, and $\ell(\gamma_{\,|[0,t]}) = t\ell(\gamma)$ by [28, Lemma 5.18]. For definitions of parallel transport and geodesic, see Appendix A.1 & A.4.

C Proofs of Section 3

C.1 Proof of Theorem 2

We begin the proof by observing the following lemma:

Lemma 11. *Assume* ***A*** *1,* ***A*** *2. Then for any $n \in \mathbb{N}^*$ and $\varepsilon > 0$,*

$$
\begin{aligned}
\sum_{k=0}^{n} \eta_{k+1}(\underline{c} - (3L/2)\eta_{k+1} - \bar{c}\varepsilon) \|h(\theta_k)\|_{\theta_k}^2 &\leq V(\theta_0) - V(\theta_{n+1}) + (3L/2)\sum_{k=0}^{n} \eta_{k+1}^2 \|e_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \\
&+ \sum_{k=0}^{n} \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\rangle_{\theta_k} + \sum_{k=0}^{n} \eta_{k+1}\left\{(4\varepsilon)^{-1} + (3L/2)\eta_{k+1}\right\} \|b_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \;,
\end{aligned}
$$

where e is defined by (5).

Proof. For any $k \geq 0$ and $t \in [0,1]$, consider $\gamma_t^{(k)} = \operatorname{Exp}_{\theta_k}\{t\eta_{k+1}(H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}))\}$. Note that $\dot{\gamma}_0^{(k)} = \eta_{k+1}\{H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\}$ and $\ell(\gamma^{(k)}) = \eta_{k+1}\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\|_{\theta_k}$. Then, by Lemma 1, (5) and using that for any $\theta \in \Theta$, $a, b, c \in \mathrm{T}_\theta\Theta$, $\|a+b+c\|_\theta^2 \leq 3(\|a\|_\theta^2 + \|b\|_\theta^2 + \|c\|_\theta^2)$, we get that for any $k \geq 0$,

$$
\begin{aligned}
\left|V(\theta_{k+1}) - V(\theta_k) - \eta_{k+1}\langle \operatorname{grad} V(\theta_k), H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\rangle_{\theta_k}\right| &\leq (L/2)\ell(\gamma^{(k)})^2 \;, \\
&= (L\eta_{k+1}^2/2)\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \;, \\
&\leq (3L/2)\eta_{k+1}^2\left\{\|h(\theta_k)\|_{\theta_k}^2 + \|b_{\theta_k}(X_{k+1})\|_{\theta_k}^2 + \|e_{\theta_k}(X_{k+1})\|_{\theta_k}^2\right\} \;.
\end{aligned}
$$

Therefore, we get that for any $k \in \mathbb{N}^*$,

$$-\eta_{k+1} \langle \operatorname{grad} V(\theta_k), h(\theta_k) \rangle_{\theta_k} \leq V(\theta_k) - V(\theta_{k+1}) + \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) \rangle_{\theta_k}$$
$$+ \eta_{k+1} \langle \operatorname{grad} V(\theta_k), b(\theta_k) \rangle_{\theta_k} + (3L/2)\eta_{k+1}^2 \left\{ \|h(\theta_k)\|_{\theta_k}^2 + \|b_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \, \|e_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \right\} .$$

By **A**2-(a) and the Cauchy-Schwarz inequality, we obtain for any $k \in \mathbb{N}^*$ and $\varepsilon > 0$,

$$\eta_{k+1}(\underline{c} - (3L/2)\eta_{k+1}) \|h(\theta_k)\|_{\theta_k}^2 \leq V(\theta_k) - V(\theta_{k+1}) + \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1}) \rangle_{\theta_k}$$
$$+ \eta_{k+1}\bar{c}^2 \varepsilon \|h(\theta_k)\|_{\theta_k}^2 + (\eta_{k+1}/4\varepsilon) \|b(\theta_k)\|_{\theta_k}^2 + (3L/2)\eta_{k+1}^2 \left\{ \|b(\theta_k)\|_{\theta_k}^2 + \|e_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \right\} .$$

Adding these inequalities from 0 to n and rearranging terms concludes the proof. □

Equipped with the above lemma, the proof for Theorem 2 proceeds as follows. Let $n \in \mathbb{N}^*$. First note that for any $k \in \mathbb{N}$, $\mathbb{E}[\langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1}) \rangle_{\theta_k}] = \mathbb{E}\left[\langle \operatorname{grad} V(\theta_k), \mathbb{E}\left[e_{\theta_k}(X_{k+1}) | \mathcal{F}_k \right] \rangle_{\theta_k} \right] = 0$, using that θ_k is $\mathcal{F}_k$-measurable and **MD**1. Therefore, taking the expectation in the inequality given by Lemma 11 and using $\|b_{\theta_k}(X_{k+1})\|_{\theta_k} \leq b_\infty$ using **MD**1, we obtain

$$\sum_{k=0}^{n} \eta_{k+1}(\underline{c} - (3L/2)\eta_{k+1} - \bar{c}\varepsilon) \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2 \right] \leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1}) \right]$$
$$+ (3L/2) \sum_{k=0}^{n} \eta_{k+1}^2 \mathbb{E}\left[\|e_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \right] + b_\infty^2 \sum_{k=0}^{n} \eta_{k+1} \left\{ (4\varepsilon)^{-1} + (3L/2)\eta_{k+1} \right\} .$$

Since for any $k \in \mathbb{N}$, $\mathbb{E}[\|e_{\theta_k}(X_{k+1})\|_{\theta_k}^2] \leq \sigma_0^2 + \sigma_1^2 \|h(\theta_k)\|_{\theta_k}^2$ using that θ_k is $\mathcal{F}_k$-measurable and **MD**1, we get

$$\sum_{k=0}^{n} \eta_{k+1}(\underline{c} - \bar{c}\varepsilon - (3L/2)\{1 + \sigma_1^2\}\eta_{k+1}) \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2 \right]$$
$$\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1}) \right] + 3L\sigma_0^2 \Gamma_{n+1}^{(2)}/2 + b_\infty^2 \left\{ \Gamma_{n+1}/(4\varepsilon) + 3L\Gamma_{n+1}^{(2)}/2 \right\} .$$

Taking $\varepsilon = \underline{c}/(2\bar{c})$ and dividing by Γ_{n+1}, we get

$$(2\Gamma_{n+1})^{-1} \sum_{k=0}^{n} \eta_{k+1}(\underline{c} - 3L\{1 + \sigma_1^2\}\eta_{k+1}) \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2 \right]$$
$$\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1}) \right] / \Gamma_{n+1} + 3L\Gamma_{n+1}^{(2)} \left\{ \sigma_0^2 + b_\infty^2 \right\} / (2\Gamma_{n+1}) + b_\infty^2 \bar{c}/(2\underline{c}) .$$

The proof is then completed using that for any $k \in \mathbb{N}^*$, $(\underline{c} - 3L(1 + \sigma_1^2)\eta_{k+1}) \geq \underline{c}/2$ and (4).

C.2 Proof of Theorem 3

Likewise in the proof for Theorem 2, we consider:

Lemma 12. *Assume* ***A*** *1-****A*** *2-****A*** *3-****MA*** *1 hold. Let* $(\eta_k)_{k\in\mathbb{N}^*}$ *be a sequence satisfying* (9). *Then for any* $n \in \mathbb{N}$,

$$\mathbb{E}\left[-\sum_{k=0}^{n} \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\rangle_{\theta_k}\right] \leq D_{\hat{e}} \sum_{k=0}^{n} \eta_{k+1}^2 \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}\right] + \tilde{C}_{\hat{e}} \Gamma_{n+1}^{(2)} + C(\eta_1) ,$$

where $\tilde{C}_{\hat{e}} = C_{\hat{e}} - (3L/2)(b_\infty^2 + e_\infty^2)$, $D_{\hat{e}}, C_{\hat{e}}$ *and* $C(\eta_1)$ *are given by* (8).

Proof. Consider the measurable function $\hat{e} : \Theta \times \mathsf{X} \to \mathrm{T}\Theta$ which satisfies **MA**1-(ii)-(iii)-(iv) and for any $k \in \mathbb{N}^*$, consider $\gamma^{(k)} : [0,1] \to \Theta$ the geodesic between θ_{k-1} and θ_k for any $k \in \mathbb{N}^*$ defined by $\gamma^{(k+1)}(t) = \operatorname{Exp}\{t\eta_{k+1}(H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}))\}$ for any $t \in [0,1]$. Note that for any $k \in \mathbb{N}^*$,

$$\ell(\gamma^{(k)}) = \eta_k \|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\|_{\theta_k} \ . \tag{47}$$

Using that the parallel transport associated with the Levi-Civita connection is a linear isometry [28, Proposition 5.5] and $(\mathrm{T}_{01}^{\gamma})^{-1} = \mathrm{T}_{10}^{\gamma}$ by uniqueness of parallel transport [28, Theorem 4.32], we obtain the following decomposition

$$\mathbb{E}\left[-\sum_{k=0}^{n} \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\rangle_{\theta_k}\right] = -\mathbb{E}\left[\sum_{i=1}^{5} A_i\right] , \tag{48}$$

where

$$\begin{aligned}
A_1 &= \sum_{k=1}^{n} \eta_{k+1} \langle \operatorname{grad} V(\theta_k), \hat{e}_{\theta_k}(X_{k+1}) - P_{\theta_k}\hat{e}_{\theta_k}(X_k)\rangle_{\theta_k} \ , \\
A_2 &= \sum_{k=1}^{n} \eta_{k+1} \left\langle \operatorname{grad} V(\theta_k), P_{\theta_k}\hat{e}_{\theta_k}(X_k) - \mathrm{T}_{01}^{\gamma^{(k)}} P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\rangle_{\theta_k} \ , \\
A_3 & \\
&= \sum_{k=1}^{n} \eta_{k+1} \left\langle \mathrm{T}_{01}^{\gamma^{(k)}}\mathrm{T}_{10}^{\gamma^{(k)}} \operatorname{grad} V(\theta_k) - \mathrm{T}_{01}^{\gamma^{(k)}}\mathrm{T}_{10}^{\gamma^{(k)}}\mathrm{T}_{01}^{\gamma^{(k)}} \operatorname{grad} V(\theta_{k-1}), \mathrm{T}_{01}^{\gamma^{(k)}}\mathrm{T}_{10}^{\gamma^{(k)}}\mathrm{T}_{01}^{\gamma^{(k)}} P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\rangle_{\theta_k} \\
&= \sum_{k=1}^{n} \eta_{k+1} \left\langle \mathrm{T}_{10}^{\gamma^{(k)}} \operatorname{grad} V(\theta_k) - \operatorname{grad} V(\theta_{k-1}), P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\rangle_{\theta_{k-1}} \ , \\
A_4 &= \sum_{k=1}^{n} (\eta_{k+1} - \eta_k) \left\langle \operatorname{grad} V(\theta_{k-1}), P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\rangle_{\theta_{k-1}} \ , \\
A_5 &= \eta_1 \langle \operatorname{grad} V(\theta_0), \hat{e}_{\theta_0}(X_1)\rangle_{\theta_0} + \eta_{n+1} \langle \operatorname{grad} V(\theta_n), \hat{e}_{\theta_n}(X_{n+1})\rangle_{\theta_n} \ .
\end{aligned} \tag{49}$$

We now bound each terms of this decomposition. First note that using **MA** 1-(i)-(iii), we get that for any $k \in \{1, \ldots, n\}$, $\mathbb{E}[\langle \operatorname{grad} V(\theta_k), \hat{e}_{\theta_k}(X_{k+1}) - P_{\theta_k}\hat{e}_{\theta_k}(X_k)\rangle_{\theta_k} [\mathcal{F}_k] = 0$ and therefore

$$\mathbb{E}[A_1] = 0 \,. \tag{50}$$

Using Cauchy-Schwarz inequality, **MA** 1-(iii)-(iv), (47) and **A** 2-(a), we get

$$\begin{aligned}
|A_2| &\leq L_{\hat{e}} \sum_{k=1}^{n} \eta_{k+1} \left\|\operatorname{grad} V(\theta_k)\right\|_{\theta_k} \ell(\gamma^{(k)}) \,, \\
&\leq L_{\hat{e}} \sum_{k=1}^{n} \eta_{k+1}\eta_k \left\|\operatorname{grad} V(\theta_k)\right\|_{\theta_k} \left\{\|H_{\theta_{k-1}}(X_k)\|_{\theta_{k-1}} + \|b_{\theta_{k-1}}(X_k)\|_{\theta_{k-1}}\right\} \,, \\
&\leq L_{\hat{e}}\bar{c} \sum_{k=1}^{n} \eta_{k+1}\eta_k \|h(\theta_k)\|_{\theta_k} \{e_\infty + \|h(\theta_{k-1})\|_{\theta_{k-1}} + \|b_{\theta_{k-1}}(X_k)\|_{\theta_{k-1}}\} \,.
\end{aligned}$$

Using $(\eta_k)_{k\in\mathbb{N}^*}$ satisfies (9), **A** 3, for any $a, b \in \mathbb{R}$, $a \leq a^2 + 1$ and $|ab| \leq (a^2 + b^2)/2$, we get

$$|A_2| \leq L_{\hat{e}}\bar{c} \left\{ \sum_{k=1}^{n} \eta_k^2 (e_\infty + b_\infty) + \sum_{k=1}^{n} \eta_k^2 \|h(\theta_k)\|_{\theta_k}^2 (1 + e_\infty + b_\infty) \right\} \,. \tag{51}$$

Using **A** 2-(b), Lemma 1, **A** 3, **MA** 1-(iii), we obtain

$$\begin{aligned}
|A_3| &\leq L \sum_{k=1}^{n} \eta_k\eta_{k+1} \left\{\|H_{\theta_{k-1}}(X_k)\|_{\theta_{k-1}} + \|b_{\theta_{k-1}}(X_k)\|_{\theta_{k-1}}\right\} \left\|P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\|_{\theta_{k-1}} \,, \\
&\leq L\hat{e}_\infty \sum_{k=1}^{n} \eta_k\eta_{k+1} \left(e_\infty + b_\infty + \|h(\theta_{k-1})\|_{\theta_{k-1}}\right) \,, \\
&\leq L\hat{e}_\infty \left\{ (e_\infty + b_\infty + 1) \sum_{k=1}^{n} \eta_k^2 + \sum_{k=1}^{n} \eta_k^2 \|h(\theta_{k-1})\|_{\theta_{k-1}}^2 \right\} \,.
\end{aligned} \tag{52}$$

Using the Cauchy-Schwarz inequality, **A** 2-(a), **MA** 1-(iii), for any $a \in \mathbb{R}$, $a \leq 1 + a^2$ and that $(\eta_k)_{k\in\mathbb{N}^*}$ satisfies (9), we have

$$\begin{aligned}
|A_4| &\leq \bar{c} \sum_{k=1}^{n} |\eta_{k+1} - \eta_k| \, \|h(\theta_{k-1})\|_{\theta_{k-1}} \left\|P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\|_{\theta_{k-1}} \,, \\
&\leq \bar{c}\hat{e}_\infty \sum_{k=1}^{n} |\eta_{k+1} - \eta_k| \, \|h(\theta_{k-1})\|_{\theta_{k-1}} \leq \bar{c}\hat{e}_\infty \left\{ \eta_1 + a_2 \sum_{k=1}^{n} \eta_k^2 \|h(\theta_{k-1})\|_{\theta_{k-1}}^2 \right\} \,.
\end{aligned} \tag{53}$$

Finally, by the Cauchy-Schwarz inequality, **A** 2-(a), **MA** 1-(iii) and using that $(\eta_k)_{k\in\mathbb{N}^*}$ is nonincreasing, we obtain that

$$\begin{aligned}
|A_5| &\leq \bar{c}\hat{e}_\infty \left\{ \eta_1 \|h(\theta_0)\|_{\theta_0} + \eta_{n+1} \|h(\theta_n)\|_{\theta_n} \right\} \,, \\
&\leq \bar{c}\hat{e}_\infty \left\{ 2 + \eta_1^2 \|h(\theta_0)\|_{\theta_0}^2 + \eta_{n+1}^2 \|h(\theta_n)\|_{\theta_n}^2 \right\} \,.
\end{aligned} \tag{54}$$

Combining (50)-(51)-(52)-(53)-(54) in (48) completes the proof. □

Proof of Theorem 3. The proof only consists in applying Lemma 11 taking $\varepsilon = \underline{c}/(2\overline{c})$, Lemma 12 and using the last inequality condition in (9).

□

D Omitted Proofs for Section 4

D.1 Proof of Lemma 4

We preface the proof of the Lemma by a preliminary result which does not assume **R**3.

Lemma 13. *Assume* ***R****1,***R****2 hold.*

(a) Under ***R****4,* $\|\Phi_\theta(u) - u\|_\theta \leq \mathscr{L}^{(1)}(\theta)\|u\|_\theta^2/2$ *for any* $(\theta, u) \in \mathrm{T}\Theta$*.*

(b) Under ***R****5,* $\|\Phi_\theta(u) - u\|_\theta \leq \mathscr{L}^{(2)}(\theta)\|u\|_\theta^3/6$ *for any* $(\theta, u) \in \mathrm{T}\Theta$*.*

Proof. (a) Let $\theta \in \Theta$ and for any $u \in \mathrm{T}_\theta\Theta$, consider the first-order Taylor expansion of $\Phi_\theta : T_\theta\Theta \to T_\theta\Theta$, taken at 0_θ,

$$\Phi_\theta(u) = \Phi_\theta(0_\theta) + \mathrm{D}\Phi_\theta(0_\theta)[u] + \int_0^1 (1-t)\,\mathrm{D}^2\Phi_\theta(tu)[u,u]\,\mathrm{d}t\ , \tag{55}$$

where $\mathrm{D}\Phi_\theta$ and $\mathrm{D}^2\Phi_\theta$ denote the first and second derivative of Φ_θ. For the first term,

$$\Phi_\theta(0_\theta) \ = \ \mathrm{Exp}_\theta^{-1} \circ \mathrm{Ret}_\theta(0_\theta) \ = \ \mathrm{Exp}_\theta^{-1}(\theta) \ = \ 0_\theta\ , \tag{56}$$

where the second equality follows because $\mathrm{Ret}_\theta(0_\theta) = \theta$, by **R**1 and by definition that $\mathrm{Exp}_\theta(0_\theta) = \theta$. For the second term, using that Exp_θ^{-1} and Ret_θ are continuously differentiable as function between smooth manifolds, we obtain since $\mathrm{DExp}_\theta(0_\theta) = \mathrm{Id}$ by definition and using **R**1 that

$$\mathrm{D}\Phi_\theta(0_\theta) = \mathrm{DExp}_\theta^{-1}(\mathrm{Ret}_\theta(0_\theta))\mathrm{DRet}_\theta(0_\theta) = \mathrm{DExp}_\theta^{-1}(\theta)\mathrm{DRet}_\theta(0_\theta) = \mathrm{Id}\ . \tag{57}$$

The proof is then completed using (56), (57) and **R**4 in (55).

(b) Let $\theta \in \Theta$ and consider the second-order Taylor expansion of $\Phi_\theta : \mathrm{T}_\theta\Theta \to \mathrm{T}_\theta\Theta$, taken at 0_θ:

$$\Phi_\theta(u) = u + \mathrm{D}^2\Phi_\theta(0_\theta)[u,u]/2 + 2^{-1}\int_0^1 (1-t)^2\mathrm{D}^3\Phi_\theta(tu)[u,u,u]\,\mathrm{d}t\ . \tag{58}$$

where $\mathrm{D}^3\Phi_\theta$ is the third derivative of Φ_θ. The proof relies on the use of normal coordinates with origin at θ [28, Chapter 5]. These coordinates are smooth and simply defined identifying $\mathrm{T}_\theta\Theta$ with $\mathbb{R}^d$ through Exp_θ^{-1}. More precisely, setting an orthonormal basis $\{\mathbf{b}_i : i \in \{1,\ldots,d\}\}$ of $\mathrm{T}_\theta\Theta$, define for any $\tilde{\theta} \notin \mathrm{Cut}(\theta)$,

$$\varphi^i(\tilde{\theta}) = \left\langle \mathrm{Exp}_\theta^{-1}(\tilde{\theta}), \mathbf{b}_i \right\rangle_\theta\ .$$

Then, $\varphi = \{\varphi^i : i \in \{1, \ldots, d\}\}$ are smooth coordinates around θ. Therefore, by definition Φ_θ is simply Ret_θ read in these coordinates. Then by [28, Equation (4.15)], setting for any $t \in \mathbb{R}_+$, $\gamma(t) = \mathrm{Ret}_\theta(tu)$ for $u \in \mathrm{T}_\theta\Theta$, we get that, in these coordinates,

$$\mathrm{D}_t\dot{\gamma}(t) = \mathrm{D}^2\Phi_\theta(tu)[u,u] + \sum_{k=1}^{d}\sum_{i,j=1}^{d} \mathrm{D}\Phi_\theta^i\dot{\gamma}^j(t)\mathbf{\Gamma}_{i,j}^k(\gamma(t))\partial_k \ ,$$

where D_t is the covariant derivative along γ, $\{\mathbf{\Gamma}_{i,j}^k : i,j,k \in \{1,\ldots,d\}\}$ are the Christoffel symbols and $\{\partial_k : k \in \{1,\ldots,d\}\}$ coordinate vector fields on $\mathrm{T}\Theta$ corresponding to φ. But using [28, Proposition 5.24], we get that $\mathbf{\Gamma}_{i,j}^k(\gamma(0)) = \mathbf{\Gamma}_{i,j}^k(\theta) = 0$ for any $i,j,k \in \{1,\ldots,d\}$. Therefore,

$$\mathrm{D}_t\dot{\gamma}(0) = \mathrm{D}^2\Phi_\theta(0)[u,u] \tag{59}$$

and by **R**5, $\mathrm{D}_t\dot{\gamma}(0) = 0$, which implies that $\mathrm{D}^2\Phi_\theta(0)[u,u] = 0_\theta$. Plugging this result into (58) and using the bound on the third derivative of Φ_θ given by **R**5 completes the proof. □

Proof of Lemma 4. (a) It will now be proven that $\mathscr{L}^{(1)}(\theta)$ can be chosen independent of θ. To do so, fix some $o \in \Theta$ and note from Lemma 13-(a),

$$\|\Phi_o(v) - v\|_o \le \mathscr{L}^{(1)}(o)\ \|v\|_o^2/2\ , \quad \text{for any } v \in \mathrm{T}_o\Theta\ . \tag{60}$$

By **R**3, Θ is a homogeneous Riemannian manifold under the action of G. Thus, for any $\theta \in \Theta$ there exists $g \in \mathsf{G}$ such that $g^{-1}\cdot\theta = o$. Taking $v = g^{-1}\cdot u$ in (60), and noting that g is an isometry (so g preserves Riemannian norms), it follows

$$\|\Phi_\theta(u) - u\|_\theta \ = \ \left\|g^{-1}\cdot(\Phi_\theta(u) - u)\right\|_o \ = \ \left\|g^{-1}\cdot\Phi_\theta(u) - v\right\|_o\ . \tag{61}$$

However, it will shortly be shown that

$$g^{-1}\cdot\Phi_\theta(u) \ = \ \Phi_o(v)\ . \tag{62}$$

Replacing this into (61),

$$\|\Phi_\theta(u) - u\|_\theta \ = \ \|\Phi_o(v) - v\|_o \le \mathscr{L}^{(1)}(o)\ \|v\|_o^2/2 \ = \ \mathscr{L}^{(1)}(o)\ \|u\|_\theta^2/2\ . \tag{63}$$

where the inequality follows from (60), and the final equality because $v = g^{-1}\cdot u$ and g is an isometry. Clearly, (63) is identical to the required inequality, with $\mathscr{L}^{(1)} = \mathscr{L}^{(1)}(o)/2$ (which is independent of θ). Now, to complete the proof it only remains to show (62). To do so, let $\rho = \mathrm{Ret}_\theta(u)$ and note that

$$g^{-1}\cdot\Phi_\theta(u) \ = \ g^{-1}\cdot\mathrm{Exp}_\theta^{-1}(\rho)\ . \tag{64}$$

Because g is an isometry, so g maps geodesics to geodesics, it follows that (see (41) in A.7)

$$g^{-1}\cdot\mathrm{Exp}_\theta^{-1}(\rho) \ = \ \mathrm{Exp}_o^{-1}(g^{-1}\cdot\rho)\ . \tag{65}$$

However, from (13) in **R**3 and the definition of ρ,

$$g^{-1} \cdot \rho \;=\; \mathrm{Ret}_o(g^{-1} \cdot u) \;=\; \mathrm{Ret}_o(v)\;.$$

Replacing this into (65), it is seen than

$$g^{-1} \cdot \mathrm{Exp}_\theta^{-1}(\rho) \;=\; \mathrm{Exp}_o^{-1}(\mathrm{Ret}_o(v)) \;=\; \Phi_o(v)\;.$$

Then, (62) follows from (64).
(b) the proof is almost identical to item (a). First, fixe some $o \in \Theta$ and note from Lemma 13-(b),

$$\|\Phi_o(v) - v\|_o \;\leq\; \mathscr{L}^{(2)}(o)\; \|v\|_o^3/6\;, \quad \text{for any } v \in T_o\Theta\;. \tag{66}$$

Then, for any $\theta \in \Theta$, write $g^{-1} \cdot \theta = o$. Taking $v = g^{-1} \cdot u$ in (66) and noting that g is an isometry, it follows

$$\|\Phi_\theta(u) - u\|_\theta \;=\; \left\|g^{-1} \cdot (\Phi_\theta(u) - u)\right\|_o \;=\; \left\|g^{-1} \cdot \Phi_\theta(u) - v\right\|_o\;.$$

Then, using (62),

$$\|\Phi_\theta(u) - u\|_\theta \;=\; \|\Phi_o(v) - v\|_o \;\leq\; \mathscr{L}^{(2)}(o)\; \|v\|_o^3/6 \;=\; \mathscr{L}^{(2)}(o)\; \|u\|_\theta^2/6\;. \tag{67}$$

where the inequality follows from (66), and the final equality because $v = g^{-1} \cdot u$ and g is an isometry. Clearly, (67) is identical to the required inequality, with $\mathscr{L}^{(1)} = \mathscr{L}^{(1)}(o)/2$ (which is independent of θ). ☐

D.2 Proof of Theorem 5

Before we begin proving Theorem 5, let us state a similar result to Theorem 5-(a), with slightly different assumptions.

Theorem 14. *Assume* ***A****1-****A****2-****A****3-****A****4,* ***R****1-****R****2-****R****3-****R****4,* ***MD****1-****MD****2(2) hold. Consider $(\theta_n)_{n\in\mathbb{N}}$ defined by (2). For any $n \in \mathbb{N}^*$ and let $I_n \in \{0, \ldots, n\}$ be distributed according to (4). Assume for any $\theta \in \Theta$, the function Φ_θ defined by (11) has first derivative bounded by $\mathscr{L}_\infty^{(0)}$. If $\sup_{k\in\mathbb{N}^*} \eta_k \leq (4L + 6\mathscr{L}_\infty^{(1)} V_\infty)^{-1}\underline{c}$, then*

$$\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2\right] \leq 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + c_0\Gamma_{n+1}^{(2)} + c_1\Gamma_{n+1}^{(4)}\right\} + 2b_\infty V_\infty/\underline{c}\;,$$

where $c_0 = 2L(5 + 8(\mathscr{L}_\infty^{(0)})^2))(\sigma_{(2)} + b_\infty^2)$, $c_1 = 8L(\mathscr{L}_\infty^{(1)})^2 h_\infty^4$.

The main idea of our proof is to exploit the geodesic expression (12) of the retraction stochastic approximation scheme. We begin by applying Lemma 1 to (12). Note that as

$$\ell(\gamma^{(k)}) = \eta_{k+1}\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}\,,$$

we have

$$\left|V(\theta_{k+1}) - V(\theta_k) - \eta_{k+1}\left\langle \operatorname{grad} V(\theta_k), H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k}\right|$$
$$\leq \left(L\eta_{k+1}^2\big/2\right)\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}^2 .$$

The above implies

$$\begin{aligned} -\eta_{k+1}\left\langle \operatorname{grad} V(\theta_k), h(\theta_k)\right\rangle_{\theta_k} \leq V(\theta_k) - V(\theta_{k+1}) &+ \eta_{k+1}\left\langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\right\rangle_{\theta_k} \\ &+ \eta_{k+1}\left\langle \operatorname{grad} V(\theta_k), b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k} \\ &+ \left(L\eta_{k+1}^2\big/2\right)\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}^2 . \end{aligned} \tag{68}$$

Applying **A**2-(a) to the left hand side yields the lower bound to the left hand side:

$$-\eta_{k+1}\left\langle \operatorname{grad} V(\theta_k), h(\theta_k)\right\rangle_{\theta_k} \geq \eta_{k+1}\underline{c}\left\|h(\theta_k)\right\|_{\theta_k}^2 .$$

Therefore taking the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_k]$ on both sides of (68) shows that:

$$\begin{aligned} &\eta_{k+1}\underline{c}\left\|h(\theta_k)\right\|_{\theta_k}^2 \\ &\quad \leq V(\theta_k) - V(\theta_{k+1}) + \eta_{k+1}\mathbb{E}\left[\left\langle \operatorname{grad} V(\theta_k), b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k}\middle|\mathcal{F}_k\right] \\ &\quad + \left(L\eta_{k+1}^2\big/2\right)\mathbb{E}\left[\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}^2\middle|\mathcal{F}_k\right] . \end{aligned} \tag{69}$$

Next, we consider different cases of retraction and prove their corresponding bounds for the last two terms in (69).

First-order Retraction Consider either case (a) in Theorem 5 or Theorem 14, where **R**4, **MD**2(2 or 4) hold. Applying the Cauchy-Schwarz inequality and **A**4 give

$$\begin{aligned} \left\langle \operatorname{grad} V(\theta_k), \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k} &\leq \left\|\operatorname{grad} V(\theta_k)\right\|_{\theta_k}\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k} , \\ &\leq V_\infty\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k} . \end{aligned}$$

Invoking Lemma 4-(a) shows that

$$\begin{aligned} \left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k} &\leq \mathscr{L}_\infty^{(1)}\eta_{k+1}\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\right\|_{\theta_k}^2 , \\ &\leq 3\mathscr{L}_\infty^{(1)}\eta_{k+1}\left\{\left\|h(\theta_k)\right\|_{\theta_k}^2 + b_\infty^2 + \left\|e_{\theta_k}(X_{k+1})\right\|_{\theta_k}^2\right\} . \end{aligned} \tag{70}$$

Therefore, the above inequalities give

$$\begin{aligned}&\eta_{k+1}\mathbb{E}\left[\left\langle \operatorname{grad} V(\theta_k), b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k}\middle|\mathcal{F}_k\right]\\&\leq \eta_{k+1}V_\infty b_\infty + 3\eta_{k+1}^2 V_\infty \mathscr{L}_\infty^{(1)}\left\{\|h(\theta_k)\|_{\theta_k}^2 + b_\infty^2 + \sigma_{(2)}\right\}\ .\end{aligned} \tag{71}$$

Next, we consider

$$\begin{aligned}&\mathbb{E}\left[\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}^2\middle|\mathcal{F}_k\right]\\&\leq 4\left\{\|h(\theta_k)\|_{\theta_k}^2 + b_\infty^2 + \sigma_{(2)} + \mathbb{E}\left[\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}^2\middle|\mathcal{F}_k\right]\right\}\ .\end{aligned} \tag{72}$$

and focus on the following sub-cases with first order retraction:

*Sub-case when **MD2**(2) holds and Φ_θ has bounded first order derivative* — Let $\mathrm{D}\Phi_{\theta_k}(x)[u]$ be the first-order derivative of Φ_{θ_k} at $x \in \mathrm{T}_{\theta_k}\Theta$ in the direction of $u \in \mathrm{T}_{\theta_k}\Theta$. Observe

$$\begin{aligned}&\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\\&\quad= \eta_{k+1}^{-1}\{\Phi_{\theta_k}(\eta_{k+1}\{H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\}) - \Phi_{\theta_k}(\eta_{k+1}h(\theta_k))\}\\&\qquad+ \eta_{k+1}^{-1}\{\Phi_{\theta_k}(\eta_{k+1}h(\theta_k)) - \eta_{k+1}\{H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\}\}\ ,\\&\quad= \int_0^1 \mathrm{D}\Phi_{\theta_k}(\eta_{k+1}(h(\theta_k) + t(e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}))))[e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})]\,\mathrm{d}t\\&\qquad+ \eta_{k+1}^{-1}\{\Phi_{\theta_k}(\eta_{k+1}h(\theta_k)) - \eta_{k+1}h(\theta_k) - \eta_{k+1}\{e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\}\}\ .\end{aligned}$$

where the last equality is due to the mean value theorem. Together with Lemma 4, the above is bounded by using Hölder's inequality, and bounding each term. The integral is treated as follows

$$\begin{aligned}&\left\|\int_0^1 \mathrm{D}\Phi_{\theta_k}(\eta_{k+1}(h(\theta_k) + t(e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}))))[e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})]\,\mathrm{d}t\right\|_{\theta_k}^2\\&\quad\leq (\mathscr{L}_\infty^{(0)})^2\,\|e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\|_{\theta_k}^2 \leq 2(\mathscr{L}_\infty^{(0)})^2(\sigma_{(2)} + b_\infty^2)\ ,\end{aligned}$$

where $\mathscr{L}_\infty^{(0)}$ upper bounds the operator norm of $\mathrm{D}\Phi_{\theta_k}(x)$. Dealing with the other terms is easier, yielding

$$\mathbb{E}\left[\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k}^2 |\mathcal{F}_k\right] \leq 4\Big\{(1 + 2(\mathscr{L}_\infty^{(0)})^2)(\sigma_{(2)} + b_\infty^2) + \eta_{k+1}^2(\mathscr{L}_\infty^{(1)})^2 h_\infty^4\Big\}\ .$$

Substituting the above into (72), and combining it with (71) into (69) leads to:

$$\begin{aligned}&\eta_{k+1}(\underline{c} - (2L + 3\mathscr{L}_\infty^{(1)}V_\infty)\eta_{k+1})\,\|h(\theta_k)\|_{\theta_k}^2\\&\leq V(\theta_k) - \mathbb{E}\left[V(\theta_{k+1})|\mathcal{F}_k\right] + \eta_{k+1}V_\infty b_\infty + \eta_{k+1}^2\{(3\mathscr{L}_\infty^{(1)}V_\infty + 2L(5 + 8(\mathscr{L}_\infty^{(0)})^2))(\sigma_{(2)} + b_\infty^2)\}\\&\quad+ \eta_{k+1}^4 8L(\mathscr{L}_\infty^{(1)})^2 h_\infty^4\ .\end{aligned}$$

As the step size is chosen such that

$$\underline{c} - (2L + 3\mathscr{L}_\infty^{(1)} V_\infty)\eta_{k+1} \geq \underline{c}/2 \,,$$

summing up the above inequality from $k = 0$ to $k = n$ gives:

$$\begin{aligned}
&(\underline{c}/2) \sum_{k=0}^{n} \eta_{k+1} \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2 \right] \\
&\leq \mathbb{E}\left[V(\theta_0)\right] + V_\infty b_\infty \Gamma_{n+1} + \{(3\mathscr{L}_\infty^{(1)} V_\infty + 2L(5 + 8(\mathscr{L}_\infty^{(0)})^2))(\sigma_{(2)} + b_\infty^2)\}\Gamma_{n+1}^{(2)} + 8L(\mathscr{L}_\infty^{(1)})^2 h_\infty^4 \Gamma_{n+1}^{(4)} \,.
\end{aligned}$$

Using that $V(\theta_{n+1}) \geq 0$ yields

$$\begin{aligned}
&\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2 \right] \\
&\leq (2/\underline{c}) \Big\{ \mathbb{E}\left[V(\theta_0)\right]/\Gamma_{n+1} + \{(3\mathscr{L}_\infty^{(1)} V_\infty + 2L(5 + 8(\mathscr{L}_\infty^{(0)})^2))(\sigma_{(2)} + b_\infty^2)\} \, \Gamma_{n+1}^{(2)} \Big/ \Gamma_{n+1} \\
&\qquad + 8L(\mathscr{L}_\infty^{(1)})^2 h_\infty^4 \Gamma_{n+1}^{(4)} \Big/ \Gamma_{n+1} + V_\infty b_\infty \Big\} \,.
\end{aligned}$$

The above inequality yields the desired bound in Theorem 14.

Sub-case when ***MD*** *2(4) holds* — invoking Lemma 4-(a) we observe that

$$\begin{aligned}
\mathbb{E}\left[\left\| \Delta_{\theta_k, \eta_{k+1}}(X_{k+1}) \right\|_{\theta_k}^2 \middle| \mathcal{F}_k \right] &\leq \eta_{k+1}^2 (\mathscr{L}_\infty^{(1)})^2 \mathbb{E}\left[\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\|_{\theta_k}^4 \middle| \mathcal{F}_k \right] \,, \\
&\leq 27\eta_{k+1}^2 (\mathscr{L}_\infty^{(1)})^2 \{ h_\infty^4 + b_\infty^4 + \sigma_{(4)} \} \,.
\end{aligned}$$

where we have used $H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) = h(\theta_k) + e_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})$ and **A**4 in the last inequality. Substituting the above into (69) leads to:

$$\begin{aligned}
&\eta_{k+1}(\underline{c} - (2L + 3\mathscr{L}_\infty^{(1)} V_\infty)\eta_{k+1}) \, \|h(\theta_k)\|_{\theta_k}^2 \\
&\leq V(\theta_k) - \mathbb{E}\left[V(\theta_{k+1}) | \mathcal{F}_k\right] + \eta_{k+1} V_\infty b_\infty + \eta_{k+1}^2 (3\mathscr{L}_\infty^{(1)} V_\infty + 2L)\{b_\infty^2 + \sigma_{(2)}\} \\
&\quad + 54L(\mathscr{L}_\infty^{(1)})^2 \eta_{k+1}^4 \{h_\infty^4 + b_\infty^4 + \sigma_{(4)}\} \,.
\end{aligned}$$

As we have set

$$\underline{c} - (2L + 3\mathscr{L}_\infty^{(1)} V_\infty)\eta_{k+1} \geq \underline{c}/2 \,,$$

taking the full expectation and summing up the previous inequality from $k = 0$ to $k = n$ leads to

$$\begin{aligned}
&(\underline{c}/2) \sum_{k=0}^{n} \eta_{k+1} \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2 \right] \\
&\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1})\right] + V_\infty b_\infty \Gamma_{n+1} + \{(3\mathscr{L}_\infty^{(1)} V_\infty + 2L)(\sigma_{(2)} + b_\infty^2)\}\Gamma_{n+1}^{(2)} \\
&\quad + 54L(\mathscr{L}_\infty^{(1)})^2 \{h_\infty^4 + b_\infty^4 + \sigma_{(4)}\}\Gamma_{n+1}^{(4)} \,.
\end{aligned}$$

Hence, using that $V(\theta_{n+1}) \geq 0$ for any $n \in \mathbb{N}$,

$$
\begin{aligned}
&(\underline{c}/2)\mathbb{E}\left[\|h(\theta_{I_n})\|^2_{\theta_{I_n}}\right] \\
&\leq \mathbb{E}\left[V(\theta_0)\right]/\Gamma_{n+1} + (3\mathscr{L}^{(1)}_\infty V_\infty + 2L)\{\sigma_{(2)} + b^2_\infty\}\, \Gamma^{(2)}_{n+1}\Big/\Gamma_{n+1} \\
&\quad + 54L(\mathscr{L}^{(1)}_\infty)^2\{h^4_\infty + b^4_\infty + \sigma_{(4)}\}\, \Gamma^{(4)}_{n+1}\Big/\Gamma_{n+1} + V_\infty b_\infty \ .
\end{aligned}
$$

The above inequality yields the desirable bound in case (a).

Second-order Retraction We consider the case where **R** 5, **MD** 2(6) hold. Invoking Lemma 4-(b), we have that

$$
\begin{aligned}
\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|_{\theta_k} &\leq \eta^2_{k+1}\mathscr{L}^{(2)}_\infty \left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\right\|^3_{\theta_k} \ , \\
&\leq 9\eta^2_{k+1}\mathscr{L}^{(2)}_\infty\{\ \|e_{\theta_k}(X_{k+1})\|^3_{\theta_k} + h^3_\infty + b^3_\infty\} \ .
\end{aligned}
\tag{73}
$$

Thus,

$$
\mathbb{E}\left[\eta_{k+1}\left\langle \operatorname{grad} V(\theta_k), \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k} |\mathcal{F}_k\right] \leq 9\mathscr{L}^{(2)}_\infty V_\infty \eta^3_{k+1}\{h^3_\infty + \sigma_{(3)} + b^3_\infty\} \ .
$$

Moreover, since

$$
\begin{aligned}
\mathbb{E}\left[\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|^2_{\theta_k} |\mathcal{F}_k\right] &\leq \eta^4_{k+1}(\mathscr{L}^{(2)}_\infty)^2\mathbb{E}\left[\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1})\|^6_{\theta_k} |\mathcal{F}_k\right] \ , \\
&\leq 3^5\eta^4_{k+1}(\mathscr{L}^{(2)}_\infty)^2\{h^6_\infty + b^6_\infty + \sigma_{(6)}\} \ ,
\end{aligned}
$$

we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|^2_{\theta_k} |\mathcal{F}_k\right] \\
&\leq 4\Big\{\ \|h(\theta_k)\|^2_{\theta_k} + b^2_\infty + \sigma_{(2)} + \mathbb{E}\left[\left\|\Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|^2_{\theta_k} |\mathcal{F}_k\right]\Big\} \ , \\
&\leq 4\Big\{\ \|h(\theta_k)\|^2_{\theta_k} + b^2_\infty + \sigma_{(2)} + 243(\mathscr{L}^{(2)}_\infty)^2\eta^4_{k+1}(h^6_\infty + b^6_\infty + \sigma_{(6)})\Big\} \ .
\end{aligned}
$$

Substituting the above into (69) leads to:

$$
\begin{aligned}
&\eta_{k+1}\{\underline{c} - 2L\eta_{k+1}\}\,\|h(\theta_k)\|^2_{\theta_k} \\
&\leq V(\theta_k) - \mathbb{E}\left[V(\theta_{k+1})|\mathcal{F}_k\right] + \eta_{k+1}V_\infty b_\infty + 9\mathscr{L}^{(2)}_\infty V_\infty \eta^3_{k+1}\{h^3_\infty + \sigma_{(3)} + b^3_\infty\} \\
&\quad + 2L\eta^2_{k+1}\Big\{b^2_\infty + \sigma_{(2)} + 3^5(\mathscr{L}^{(2)}_\infty)^2\eta^4_{k+1}(h^6_\infty + b^6_\infty + \sigma_{(6)})\Big\} \ .
\end{aligned}
$$

As we have set

$$
\underline{c} - 2L\eta_{k+1} \geq (\underline{c}/2) \ ,
$$

taking the full expectation and summing up both sides of the previous inequality from $k = 0$ to $k = n$ yields

$$
\begin{aligned}
&(\underline{c}/2) \sum_{k=0}^{n} \eta_{k+1} \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}\right] \\
&\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1})\right] + V_\infty b_\infty \Gamma_{n+1} + 2L\{b_\infty^2 + \sigma_{(2)}\}\Gamma_{n+1}^{(2)} \\
&\quad + 9\mathscr{L}_\infty^{(2)} V_\infty \{h_\infty^3 + \sigma_{(3)} + b_\infty^3\}\Gamma_{n+1}^{(3)} + 486L(\mathscr{L}_\infty^{(2)})^2(h_\infty^6 + b_\infty^6 + \sigma_{(6)})\Gamma_{n+1}^{(6)} .
\end{aligned}
$$

As such,

$$
\begin{aligned}
&(\underline{c}/2)\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0)\right]/\Gamma_{n+1} + 2L\{b_\infty^2 + \sigma_{(2)}\}\ \Gamma_{n+1}^{(2)}\Big/\Gamma_{n+1} + 9\mathscr{L}_\infty^{(2)} V_\infty \{h_\infty^3 + \sigma_{(3)} + b_\infty^3\}\ \Gamma_{n+1}^{(3)}\Big/\Gamma_{n+1} \\
&\quad + 486L(\mathscr{L}_\infty^{(2)})^2\{h_\infty^6 + b_\infty^6 + \sigma_{(6)}\}\ \Gamma_{n+1}^{(6)}\Big/\Gamma_{n+1} + V_\infty b_\infty .
\end{aligned}
$$

The above inequality yields the desirable bound in case (b).

D.3 Proof of Theorem 6

We begin the proof by showing a similar lemma to Lemma 12 as follows:

Lemma 15. *Assume **A1**-**A2**-**A3**-**A4**-**MA1**. Let $(\eta_k)_{k\in\mathbb{N}^*}$ be a sequence satisfying (14). We have defined*

$$
\begin{aligned}
&\widetilde{C}_{\hat{e}}^{\mathrm{Ret}} = \{L_{\hat{e}}\bar{c} + L\hat{e}_\infty\}(e_\infty + b_\infty + 1) , \quad C^{\mathrm{Ret}}(\eta_1) = \bar{c}\hat{e}_\infty(2 + \eta_1) , \\
&D^{\mathrm{Ret}} = \bar{c}\hat{e}_\infty(a_2 + 1) + L_{\hat{e}}\bar{c}(a_1(b_\infty + e_\infty) + 1) + L\hat{e}_\infty , \\
&P_n := D^{\mathrm{Ret}} \sum_{k=0}^{n} \eta_{k+1}^2 \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] + \widetilde{C}_{\hat{e}}^{\mathrm{Ret}}\Gamma_{n+1}^{(2)} + C^{\mathrm{Ret}}(\eta_1) .
\end{aligned}
$$

Note that $\widetilde{C}_{\hat{e}}^{\mathrm{Ret}} = C_{\hat{e}}^{\mathrm{Ret}} - 2L(b_\infty^2 + e_\infty^2)$.

*(a) Under **R4**, it holds that for any $n \in \mathbb{N}$,*

$$
\mathbb{E}\left[-\sum_{k=0}^{n} \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\rangle_{\theta_k}\right] \leq P_n + E_{(1)}^{\mathrm{Ret}}\Gamma_{n+1}^{(3)} , \tag{74}
$$

where $E_{(1)}^{\mathrm{Ret}} = 3\hat{e}_\infty\mathscr{L}_\infty^{(1)}(L_{\hat{e}}h_\infty + L)(e_\infty^2 + b_\infty^2 + h_\infty^2)$.

*(b) Under **R5**, it holds that for any $n \in \mathbb{N}$,*

$$
\mathbb{E}\left[-\sum_{k=0}^{n} \eta_{k+1} \langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\rangle_{\theta_k}\right] \leq P_n + E_{(2)}^{\mathrm{Ret}}\Gamma_{n+1}^{(4)} , \tag{75}
$$

where $E_{(2)}^{\mathrm{Ret}} = 9\hat{e}_\infty\mathscr{L}_\infty^{(2)}(L_{\hat{e}}h_\infty + L)(e_\infty^3 + b_\infty^3 + h_\infty^3)$.

Proof. The proof is almost identical to that of Lemma 12 with the following modifications: first, we observe that (47) is updated to

$$\begin{aligned}\ell(\gamma^{(k)}) &= \eta_k \left\| H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1}) \right\|_{\theta_k} , \\ &\leq \eta_k \left\{ \| H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) \|_{\theta_k} + \left\| \Delta_{\theta_k,\eta_{k+1}}(X_{k+1}) \right\|_{\theta_k} \right\} .\end{aligned} \tag{76}$$

Next we can consider the same decomposition of the left hand side of (48):

$$\mathbb{E}\left[-\sum_{k=0}^{n} \eta_{k+1} \left\langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1}) \right\rangle_{\theta_k} \right] = \mathbb{E}\left[-\sum_{i=1}^{5} A_i \right] ,$$

where A_i are defined to be exactly the same as (49). In particular, we have $\mathbb{E}[A_1] = 0$ and

$$\begin{aligned}|A_4| &\leq \bar{c}\hat{e}_\infty \left\{ \eta_1 + a_2 \sum_{k=1}^{n} \eta_k^2 \| h(\theta_{k-1}) \|_{\theta_{k-1}}^2 \right\} , \\ |A_5| &\leq \bar{c}\hat{e}_\infty \left\{ 2 + \eta_1^2 \| h(\theta_0) \|_{\theta_0}^2 + \eta_{n+1}^2 \| h(\theta_n) \|_{\theta_n}^2 \right\} .\end{aligned}$$

Our remaining task is to bound $|A_2|, |A_3|$ using the updated bound (76). Observe that using Cauchy-Schwarz inequality, **MA**1-(iii)-(iv), (76) and **A**2-(a), we get

$$\begin{aligned}|A_2| &\leq L_{\hat{e}} \sum_{k=1}^{n} \eta_{k+1} \| \operatorname{grad} V(\theta_k) \|_{\theta_k} \ell(\gamma^{(k)}) , \\ &\leq L_{\hat{e}} \sum_{k=1}^{n} \eta_{k+1}\eta_k \| \operatorname{grad} V(\theta_k) \|_{\theta_k} \left\{ \| H_{\theta_{k-1}}(X_k) \|_{\theta_{k-1}} + \| b_{\theta_{k-1}}(X_k) \|_{\theta_{k-1}} + \| \Delta_{\theta_{k-1},\eta_k}(X_k) \|_{\theta_{k-1}} \right\} , \\ &\leq L_{\hat{e}}\bar{c} \sum_{k=1}^{n} \eta_{k+1}\eta_k \| h(\theta_k) \|_{\theta_k} \left\{ e_\infty + b_\infty + \| h(\theta_{k-1}) \|_{\theta_{k-1}} + \| \Delta_{\theta_{k-1},\eta_k}(X_k) \|_{\theta_{k-1}} \right\} .\end{aligned}$$

Furthermore, using the inequality $a \leq a^2 + 1$ and $|ab| \leq a^2/2 + b^2/2$ gives:

$$|A_2| \leq L_{\hat{e}}\bar{c} \left\{ (b_\infty + e_\infty)\Gamma_{n+1}^{(2)} + (a_1(b_\infty + e_\infty) + 1) \sum_{k=0}^{n} \eta_{k+1}^2 \| h(\theta_k) \|_{\theta_k}^2 + \sum_{k=1}^{n} \eta_k^2 h_\infty \| \Delta_{\theta_{k-1},\eta_k}(X_k) \|_{\theta_{k-1}} \right\} .$$

For <u>*first order retraction*</u>, i.e., under **R**4, using (14), (70), **A**4, we have

$$\begin{aligned}|A_2| &\leq L_{\hat{e}}\bar{c} \left\{ (b_\infty + e_\infty)\Gamma_{n+1}^{(2)} + (a_1(b_\infty + e_\infty) + 1) \sum_{k=0}^{n} \eta_{k+1}^2 \| h(\theta_k) \|_{\theta_k}^2 \right\} \\ &\quad + 3L_{\hat{e}}\bar{c}\mathscr{L}_\infty^{(1)} \sum_{k=1}^{n} \eta_k^3 h_\infty \left\{ b_\infty^2 + e_\infty^2 + \| h(\theta_{k-1}) \|_{\theta_{k-1}}^2 \right\} , \\ &\leq L_{\hat{e}}\bar{c} \left\{ (b_\infty + e_\infty)\Gamma_{n+1}^{(2)} + (a_1(b_\infty + e_\infty) + 1) \sum_{k=0}^{n} \eta_{k+1}^2 \| h(\theta_k) \|_{\theta_k}^2 + 3\mathscr{L}_\infty^{(1)} h_\infty (b_\infty^2 + e_\infty^2 + h_\infty^2)\Gamma_{n+1}^{(3)} \right\} .\end{aligned}$$

For *second order retraction*, i.e., under **R**5, similarly we can use (73) to obtain:

$$|A_2| \leq L_{\hat{e}}\bar{c}\left\{(b_\infty + e_\infty)\Gamma^{(2)}_{n+1} + (a_1(b_\infty + e_\infty) + 1)\sum_{k=0}^{n} \eta_{k+1}^2 \|h(\theta_k)\|^2_{\theta_k} + 9\mathscr{L}^{(2)} h_\infty (b_\infty^3 + e_\infty^3 + h_\infty^3)\Gamma^{(4)}_{n+1}\right\} .$$

On the other hand, we have

$$\begin{aligned}|A_3| &\leq L\sum_{k=1}^{n} \eta_k \eta_{k+1}\left\{\|H_{\theta_{k-1}}(X_k) + b_{\theta_{k-1}}(X_k) + \Delta_{\theta_{k-1},\eta_k}(X_k)\|_{\theta_{k-1}}\right\}\left\|P_{\theta_{k-1}}\hat{e}_{\theta_{k-1}}(X_k)\right\|_{\theta_{k-1}} , \\ &\leq L\hat{e}_\infty \sum_{k=1}^{n} \eta_k^2 \left(e_\infty + b_\infty + 1 + \|h(\theta_{k-1})\|^2_{\theta_{k-1}} + \|\Delta_{\theta_{k-1},\eta_k}(X_k)\|_{\theta_{k-1}}\right) .\end{aligned}$$

For *first order retraction*, i.e., under **R**4, again using (70), **A**4, we have

$$|A_3| \leq L\hat{e}_\infty \left\{(e_\infty + b_\infty + 1)\Gamma^{(2)}_{n+1} + \sum_{k=0}^{n} \eta_{k+1}^2 \|h(\theta_k)\|^2_{\theta_k} + 3\mathscr{L}^{(1)}(e_\infty^2 + b_\infty^2 + h_\infty^2)\Gamma^{(3)}_{n+1}\right\} .$$

For *second order retraction*, i.e., under **R**5, using (73) yields that

$$|A_3| \leq L\hat{e}_\infty \left\{(e_\infty + b_\infty + 1)\Gamma^{(2)}_{n+1} + \sum_{k=0}^{n} \eta_{k+1}^2 \|h(\theta_k)\|^2_{\theta_k} + 9\mathscr{L}^{(2)}(e_\infty^3 + b_\infty^3 + h_\infty^3)\Gamma^{(4)}_{n+1}\right\}.$$

This concludes the proof. □

Proof of Theorem 6. Following the proof for Theorem 5 but without invoking **MD**1, we obtain the following inequality that is similar to (69):

$$\begin{aligned}\eta_{k+1}\underline{c}\|h(\theta_k)\|^2_{\theta_k} \leq V(\theta_k) - V(\theta_{k+1}) &+ \eta_{k+1}\mathbb{E}\left[\langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\rangle_{\theta_k}\middle|\mathcal{F}_k\right] \\ &+ \eta_{k+1}\mathbb{E}\left[\left\langle \operatorname{grad} V(\theta_k), b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k}\middle|\mathcal{F}_k\right] \\ &+ (L\eta_{k+1}^2/2)\mathbb{E}\left[\left\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\|^2_{\theta_k}\middle|\mathcal{F}_k\right] .\end{aligned} \tag{77}$$

First order retraction Under **R**4, the terms on right hand side in (77) can be upper bounded one-by-one as:

$$\begin{aligned}&\mathbb{E}\left[\left\langle \operatorname{grad} V(\theta_k), b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k}\right] \\ &\qquad \leq V_\infty b_\infty + 3\eta_{k+1}V_\infty \mathscr{L}^{(1)}_\infty\{\,\|h(\theta_k)\|^2_{\theta_k} + b_\infty^2 + e_\infty^2\} , \\ &\mathbb{E}\left[\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\|^2_{\theta_k}\right] \\ &\qquad \leq 4\left\{\|h(\theta_k)\|^2_{\theta_k} + b_\infty^2 + e_\infty^2 + 27(\mathscr{L}^{(1)}_\infty)^2(h_\infty^4 + b_\infty^4 + e_\infty^4)\eta_{k+1}^2\right\} .\end{aligned}$$

As such, summing up the inequality (77) from $k=0$ to $k=n$ and rearranging terms yield

$$\begin{aligned}
&\sum_{k=0}^{n} \eta_{k+1}\{\underline{c} - (3V_\infty \mathscr{L}_\infty^{(1)} + 2L)\eta_{k+1}\}\, \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1})\right] + \mathbb{E}\left[\sum_{k=0}^{n} \eta_{k+1} \left\langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\right\rangle_{\theta_k}\right] + V_\infty b_\infty \Gamma_{n+1} \\
&\quad + \{3\mathscr{L}_\infty^{(1)} V_\infty + 2L\}(b_\infty^2 + e_\infty^2)\Gamma_{n+1}^{(2)} + 54L(\mathscr{L}_\infty^{(1)})^2(h_\infty^4 + b_\infty^4 + e_\infty^4)\Gamma_{n+1}^{(4)} \,.
\end{aligned} \tag{78}$$

Therefore, combining (78) and (74), we obtain

$$\begin{aligned}
&\sum_{k=0}^{n} \eta_{k+1}\{\underline{c} - (3V_\infty \mathscr{L}_\infty^{(1)} + 2L + D^{\mathrm{Ret}})\eta_{k+1}\}\mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1})\right] + C^{\mathrm{Ret}}(\eta_1) + C_{\hat{e}}^{\mathrm{Ret}}\Gamma_{n+1}^{(2)} + E_{(1)}^{\mathrm{Ret}}\Gamma_{n+1}^{(3)} + V_\infty b_\infty \Gamma_{n+1} \\
&\quad + \{3\mathscr{L}_\infty^{(1)} V_\infty + 2L\}(b_\infty^2 + e_\infty^2)\Gamma_{n+1}^{(2)} + 54L(\mathscr{L}_\infty^{(1)})^2(h_\infty^4 + b_\infty^4 + e_\infty^4)\Gamma_{n+1}^{(4)} \,.
\end{aligned}$$

Since $\eta_{k+1} \leq \underline{c}/(2(3V_\infty \mathscr{L}_\infty^{(1)} + 2L + D^{\mathrm{Ret}}))$, using $V(\theta) \geq 0$ gives

$$\begin{aligned}
&(\underline{c}/2)\sum_{k=0}^{n} \eta_{k+1}\mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0)\right] + C^{\mathrm{Ret}}(\eta_1) + V_\infty b_\infty \Gamma_{n+1} + \left\{C_{\hat{e}}^{\mathrm{Ret}} + (3\mathscr{L}_\infty^{(1)} V_\infty + 2L)(b_\infty^2 + e_\infty^2)\right\}\Gamma_{n+1}^{(2)} \\
&\quad + E_{(1)}^{\mathrm{Ret}}\Gamma_{n+1}^{(3)} + 54L(\mathscr{L}_\infty^{(1)})^2(h_\infty^4 + b_\infty^4 + e_\infty^4)\Gamma_{n+1}^{(4)} \,.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2\right] \\
&\leq 2V_\infty b_\infty/\underline{c} + 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + C^{\mathrm{Ret}}(\eta_1) + \left\{C_{\hat{e}}^{\mathrm{Ret}} + (3\mathscr{L}_\infty^{(1)} V_\infty + 2L)(b_\infty^2 + e_\infty^2)\right\}\Gamma_{n+1}^{(2)}\right\} \\
&\quad + 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{E_{(1)}^{\mathrm{Ret}}\Gamma_{n+1}^{(3)} + 54L(\mathscr{L}_\infty^{(1)})^2(h_\infty^4 + b_\infty^4 + e_\infty^4)\Gamma_{n+1}^{(4)}\right\} \,.
\end{aligned}$$

Collecting terms and computing the constants yield (15).

Second order retraction Under **R**5, the terms on right hand side in (77) can be upper bounded one-by-one as:

$$\mathbb{E}\left[\left\langle \operatorname{grad} V(\theta_k), b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\right\rangle_{\theta_k}\right] \leq V_\infty b_\infty + 9\eta_{k+1}^2 V_\infty \mathscr{L}_\infty^{(2)}\{h_\infty^3 + b_\infty^3 + e_\infty^3\} \,,$$

$$\begin{aligned}
&\mathbb{E}\left[\|H_{\theta_k}(X_{k+1}) + b_{\theta_k}(X_{k+1}) + \Delta_{\theta_k,\eta_{k+1}}(X_{k+1})\|_{\theta_k}^2\right] \\
&\qquad \leq 4\left\{\|h(\theta_k)\|_{\theta_k}^2 + b_\infty^2 + e_\infty^2 + 243(\mathscr{L}_\infty^{(2)})^2(h_\infty^6 + b_\infty^6 + e_\infty^6)\eta_{k+1}^4\right\} \,.
\end{aligned}$$

As such, summing up the inequality (77) from $k = 0$ to $k = n$ and rearranging terms yield

$$
\begin{aligned}
&\sum_{k=0}^{n} \eta_{k+1}\{\underline{c} - 2L\eta_{k+1}\}\, \mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1})\right] + \mathbb{E}\left[\sum_{k=0}^{n} \eta_{k+1} \left\langle \operatorname{grad} V(\theta_k), e_{\theta_k}(X_{k+1})\right\rangle_{\theta_k}\right] + V_\infty b_\infty \Gamma_{n+1} \\
&\quad + 2L(b_\infty^2 + e_\infty^2)\Gamma_{n+1}^{(2)} + 9\mathscr{L}_\infty^{(2)} V_\infty (h_\infty^3 + b_\infty^3 + e_\infty^3)\Gamma_{n+1}^{(3)} + 486L(\mathscr{L}_\infty^{(2)})^2(h_\infty^6 + b_\infty^6 + e_\infty^6)\Gamma_{n+1}^{(6)} \,.
\end{aligned}
$$

Combining with (75) leads to

$$
\begin{aligned}
&\sum_{k=0}^{n} \eta_{k+1}\{\underline{c} - (2L + D^{\mathrm{Ret}})\eta_{k+1}\}\mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0) - V(\theta_{n+1})\right] + C^{\mathrm{Ret}}(\eta_1) + C_{\hat{e}}^{\mathrm{Ret}}\Gamma_{n+1}^{(2)} + E_{(2)}^{\mathrm{Ret}}\Gamma_{n+1}^{(3)} + V_\infty b_\infty \Gamma_{n+1} \\
&\quad + 2L(b_\infty^2 + e_\infty^2)\Gamma_{n+1}^{(2)} + 9\mathscr{L}_\infty^{(2)} V_\infty (h_\infty^3 + b_\infty^3 + e_\infty^3)\Gamma_{n+1}^{(3)} + 486L(\mathscr{L}_\infty^{(2)})^2(h_\infty^6 + b_\infty^6 + e_\infty^6)\Gamma_{n+1}^{(6)} \,.
\end{aligned}
$$

Since we have set $(2L + D^{\mathrm{Ret}})\eta_{k+1} \leq \underline{c}/2$, using $V(\theta) \geq 0$ gives

$$
\begin{aligned}
&(\underline{c}/2)\sum_{k=0}^{n} \eta_{k+1}\mathbb{E}\left[\|h(\theta_k)\|_{\theta_k}^2\right] \\
&\leq \mathbb{E}\left[V(\theta_0)\right] + C^{\mathrm{Ret}}(\eta_1) + \left(C_{\hat{e}}^{\mathrm{Ret}} + 2L(b_\infty^2 + e_\infty^2)\right)\Gamma_{n+1}^{(2)} + V_\infty b_\infty \Gamma_{n+1} \\
&\quad + \{E_{(2)}^{\mathrm{Ret}} + 9\mathscr{L}_\infty^{(2)} V_\infty (h_\infty^3 + b_\infty^3 + e_\infty^3)\}\Gamma_{n+1}^{(3)} + 486L(\mathscr{L}_\infty^{(2)})^2(h_\infty^6 + b_\infty^6 + e_\infty^6)\Gamma_{n+1}^{(6)} \,.
\end{aligned}
$$

Finally, it leads to

$$
\begin{aligned}
&\mathbb{E}\left[\|h(\theta_{I_n})\|_{\theta_{I_n}}^2\right] \\
&\leq 2V_\infty b_\infty/\underline{c} + 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{\mathbb{E}\left[V(\theta_0)\right] + C^{\mathrm{Ret}}(\eta_1) + \left(C_{\hat{e}}^{\mathrm{Ret}} + 2L(b_\infty^2 + e_\infty^2)\right)\Gamma_{n+1}^{(2)}\right\} \\
&\quad + 2(\underline{c}\,\Gamma_{n+1})^{-1}\left\{\{E_{(2)}^{\mathrm{Ret}} + 9\mathscr{L}_\infty^{(2)} V_\infty (h_\infty^3 + b_\infty^3 + e_\infty^3)\}\Gamma_{n+1}^{(3)} + 486L(\mathscr{L}_\infty^{(2)})^2(h_\infty^6 + b_\infty^6 + e_\infty^6)\Gamma_{n+1}^{(6)}\right\} \,.
\end{aligned}
$$

Collecting terms and identifying the constants lead to (16). □

D.4 Proof of Proposition 7

First, for any $\theta \in \mathbb{S}^d$, the Riemannian exponential map is given for any $u \in \mathrm{T}_\theta\mathbb{S}^d$ by (see [28, Proposition 5.27 and its proof])

$$
\mathrm{Exp}_\theta(u) = \cos(\|u\|)\,\theta + \sin(\|u\|)(u/\|u\|) \,. \tag{79}
$$

In addition, the retraction Ret given by (18) can be written as

$$
\mathrm{Ret}_\theta(u) \;=\; \mathrm{Exp}_\theta\left(\arctan(\|u\|)\frac{u}{\|u\|}\right) \tag{80}
$$

This can be proven by replacing the identities,

$$\cos(\arctan(\|u\|)) = \frac{1}{\sqrt{1+\|u\|^2}} \qquad \sin(\arctan(\|u\|)) = \frac{\|u\|}{\sqrt{1+\|u\|^2}}$$

into (79). Indeed, this yields,

$$\mathrm{Exp}_\theta\left(\arctan(\|u\|)\frac{u}{\|u\|}\right) \;=\; \frac{1}{\sqrt{1+\|u\|^2}}\,\theta + \frac{1}{\sqrt{1+\|u\|^2}}\,u$$

To see that this is equal to $\mathrm{Ret}_\theta(u)$, note that $1+\|u\|^2 = \|\theta+u\|^2$, because $\|\theta\| = 1$ and u is orthogonal to θ (since $u \in T_\theta\Theta$). Then, (80) follows from (18).

The following are now proven.

• Condition **R**1 is satisfied: this condition is just the definition of a retraction, as given in [2].

• Condition **R**2 is satisfied: the cut locus of a point θ on the sphere S^d is $\mathrm{Cut}(\theta) = \{-\theta\}$ [28] (Page 308). The Riemannian (that is, spherical) distance between θ and $-\theta$ is $\rho_\Theta(\theta, -\theta) = \pi$. On the other hand, from(80), $\rho_\Theta(\theta, \mathrm{Ret}_\theta(u)) < \frac{\pi}{2}$ because $\arctan(\|u\|) < \pi/2$ for all $u \in T_\theta\Theta$. It is then clear that $\mathrm{Ret}_\theta(u) \neq \{-\theta\}$ for any $u \in T_\theta\Theta$.

• Condition **R**3 is satisfied: the isometry group of $\Theta = S^d$ is $G = O(d)$, the group of $d \times d$ orthogonal matrices. The action of G on Θ is given by matrix-vector multiplication, $g\cdot\theta = g\theta$ and $g \cdot u = gu$. From (18),

$$g\cdot \mathrm{Ret}_\theta(u) = \frac{g\cdot(\theta+u)}{\|\theta+u\|} \tag{81}$$

However, since g is an orthogonal matrix, g preserves Euclidean norms, so $\|\theta + u\| = \|g\cdot(\theta+u)\|$. Replacing into (81),

$$g\cdot \mathrm{Ret}_\theta(u) = \frac{g\cdot(\theta+u)}{\|g\cdot(\theta+u)\|} \;=\; \frac{g\cdot\theta + g\cdot u}{\|g\cdot\theta + g\cdot u\|} \tag{82}$$

where the second equality follows since the action of g is linear. Finally, the right-hand side of (82) is $\mathrm{Ret}_{g\cdot\theta}(g\cdot u)$.

• Condition **R**4 is satisfied: from (80) and **R**2,

$$\Phi_\theta(u) \;=\; \arctan(\|u\|)\frac{u}{\|u\|} \tag{83}$$

The required second derivative can now be computed, thanks to the identity,

$$\mathrm{D}^2\Phi_\theta(tu)[u,u] \;=\; \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Phi_\theta(tu) \tag{84}$$

Indeed, using (83) and (84),

$$\mathrm{D}^2\Phi_\theta(tu)[u,u] \;=\; \frac{\mathrm{d}^2}{\mathrm{d}t^2}\,\arctan(t\|u\|)\frac{u}{\|u\|} \;=\; \|u\|^2\left(f_2(t\|u\|)\frac{u}{\|u\|}\right)$$

where f_2 is the second derivative of the arctan function, so $|f_2(x)| \leq 1$ for real x. Now, since $\Theta = S^d$, here $\|u\|_\theta = \|u\|$. Thus, Condition **R**4 is satisfied with $\mathscr{L}^{(1)}(\theta) = 1$.

• Condition **R**5-(i) is satisfied : recall (59) from the proof of Lemma 13. This states,

$$\mathrm{D}_t\dot{\gamma}(0) = \mathrm{D}^2\Phi_\theta(0)[u,u]$$

From (84), it then follows,

$$\mathrm{D}_t\dot{\gamma}(0) \;=\; \left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0}\Phi_\theta(tu)$$

Since Φ_θ is given by (83), an elementary calculation shows the right-hand side is here equal to zero.

• Condition **R**5-(ii) is satisfied : the proof is similar to the above one for **R**4. Here, instead of (84), it is enough to use

$$\mathrm{D}^3\Phi_\theta(tu)[u,u,u] \;=\; \frac{\mathrm{d}^3}{\mathrm{d}t^3}\Phi_\theta(tu)$$

Using (83), this shows that **R**5-(i) is satisfied with $\mathscr{L}^{(2)}(\theta) = 2$.

• Φ_θ has bounded first derivative : from (83), by differentiating,

$$\mathrm{D}\Phi_\theta(u)[v] \;=\; \frac{1}{1+\|u\|^2}\,\frac{\langle u,v\rangle}{\|u\|}\,\frac{u}{\|u\|} \;+\; \frac{\arctan(\|u\|)}{\|u\|}\left(v - \frac{\langle u,v\rangle}{\|u\|}\,\frac{u}{\|u\|}\right)$$

for any u and v in $T_\theta\Theta$. Then, by an elementary calculation, and recalling that, since $\Theta = S^d$, Riemannian scalar products and norms are equal to Euclidean ones, $\|\mathrm{D}\Phi_\theta(u)[v]\|_\theta \leq 2\,\|v\|_\theta$. Thus, the operator norm of $\mathrm{D}\Phi_\Theta(u)$ is bounded by $\overline{\mathrm{D}} = 2$.

D.5 Proof of Proposition 8

First, by [16, Equation 2.32], the Riemannian exponential map at θ is given for $u \in \mathrm{T}_\theta\mathrm{Gr}_r(\mathbb{R}^d)$, with $u = B_\perp C$, $C \in \mathbb{R}^{(d-r)\times r}$:

$$\mathrm{Exp}_\theta(u) = \left[(B\,, B_\perp)\,\exp\begin{pmatrix} 0 & -C^\top \\ C & 0 \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ 0_{d-r\times r}\end{pmatrix}\right] \tag{85}$$

where exp is the matrix exponential. In addition, we show below that the retraction Ret defined by (21) can be written on the form

$$\mathrm{Ret}_\theta(u) \;=\; \mathrm{Exp}_\theta(\Phi_\theta(u)) \qquad \Phi_\theta(u) \;=\; B_\perp\,V\arctan(a)\,U^\top \tag{86}$$

for $u \in T_\theta\Theta$ with $u = B_\perp C$ as in (19), where C has singular value decomposition $C = V\,a\,U^\top$. Here, V is $(d-r)\times(d-r)$ orthogonal and U is $r \times r$ orthogonal. Moreover, $\arctan(a)$ is obtained by taking the arctangent of each element of the matrix a. Accepting (86), it is possible to show that.

• Condition **R** 1 is satisfied : this condition is just the definition of a retraction, as given in [2].

• Condition **R** 2 is satisfied : when $\Theta = \mathrm{Gr}_r(\mathbb{R}^d)$, the cut locus of each $\theta \in \Theta$ is given by [34],

$$\mathrm{Cut}(\theta) \;=\; \left\{ \mathrm{Exp}_\theta(B_\perp C) \middle| C = V\,a\,U^\top\,;\; \|a\|_\infty \;=\; \frac{\pi}{2}\right\}$$

where $\|a\|_\infty = \max_{ij}|a_{ij}|$. From (86), for any $u \in T_\theta\Theta$, one has

$$\mathrm{Ret}_\theta(u) \;=\; \mathrm{Exp}_\theta(B_\perp C(u)) \text{ where } C(u) = V\arctan(a)\,U^\top$$

Since $\|\arctan(a)\|_\infty < \pi/2$, it follows that $\mathrm{Ret}_\theta(u) \notin \mathrm{Cut}(\theta)$.

• Condition **R** 3 is satisfied : the isometry group of $\Theta = \mathrm{Gr}_r(\mathbb{R}^d)$ is $G = O(d)$, the group of $d \times d$ orthogonal matrices. The action of G on Θ is given by $g\cdot\theta = g(\theta)$ (the image of the subspace θ of $\mathbb{R}^d$ by the orthogonal tranformation g). If $u \in T_\theta\Theta$ is given by (19), then $g \cdot u = gu$ is a matrix product.

Note that, if $\theta = [B]$ for some $B \in \mathrm{St}_r(\mathbb{R}^d)$, then $g\cdot\theta = [gB]$. Applying this property in (21),

$$g \cdot \mathrm{Ret}_\theta(u) \;=\; g\left([B+u]\right) \;=\; [gB + gu] \tag{87}$$

But, since $g\cdot\theta = [gB]$ and $gu = g\cdot u$, (87) implies

$$g\cdot\mathrm{Ret}_\theta(u) \;=\; \mathrm{Ret}_{g\cdot\theta}(g\cdot u)$$

which is (13).

• Condition **R** 4 is satisfied : the required second derivative is computed using the identity (this is a repetition of (84)),

$$\mathrm{D}^2\Phi_\theta(tu)[u,u] \;=\; \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Phi_\theta(tu) \tag{88}$$

Using (86) and (88),

$$\mathrm{D}^2\Phi_\theta(tu)[u,u] \;=\; B_\perp\,V\,(a\odot a\odot f_2(ta))\,U^\top \tag{89}$$

where $\odot$ denotes the Kronecker product, and f_2 is the second derivative of the arctan function (again, this is applied to each element of the matrix (ta)). From [16] (Page 314)

$$\left\|\mathrm{D}^2\Phi_\theta(tu)[u,u]\right\|^2_\theta \;=\; \mathrm{tr}\left((a\odot a\odot f_2(ta))(a\odot a\odot f_2(ta))^\top\right)$$

where tr denotes the trace. Then, using the fact that $|f_2(x)| \leq 1$ for real x, the right-hand side is less than $\mathrm{tr}(aa^\top)$, which is equal to $\|u\|^2_\theta$. Thus, Condition **R**4 is satisfied with $\mathscr{L}^{(1)}(\theta) = 1$.

• Condition **R**5-(i) is satisfied: recall (59) from the proof of Lemma 13. This states,

$$\mathrm{D}_t\dot{\gamma}(0) = \mathrm{D}^2\Phi_\theta(0)[u,u]$$

Setting $t = 0$ in (89), it then follows

$$\mathrm{D}_t\dot{\gamma}(0) \;=\; B_\perp\, V\,(a \odot a \odot f_2(0))\, U^\top$$

which is equal to zero since $f_2(0) = 0$.

• Condition **R**5-(ii) is satisfied: the proof is similar to the above one for **R**4. Here, instead of (88), it is enough to use

$$\mathrm{D}^3\Phi_\theta(tu)[u,u,u] \;=\; \frac{\mathrm{d}^3}{\mathrm{d}t^3}\Phi_\theta(tu)$$

by computing the derivative, as in (89), it can be shown that **R**5 is satisfied with $\mathscr{L}^{(2)}(\theta) = 2$.

• Φ_θ does not have bounded first derivative: assume $r > 1$. Recall that $\Phi_\theta(u)$ is given by (86), which can be written

$$\Phi_\theta(u) = B_\perp\varphi(\psi_\theta(u)) \tag{90}$$

where, $\psi_\theta : \mathbb{R}^{d\times r} \to \mathbb{R}^{(d-r)\times r}$ and $\varphi : \mathbb{R}^{(d-r)\times r} \to \mathbb{R}^{(d-r)\times r}$ are given by

$$\psi_\theta(u) = B_\perp^\top u \qquad \varphi(C) = V\,\arctan(a)\,U^\top \tag{91}$$

whenever C has singular value decomposition $C = V\,a\,U^\top$. Indeed, if $u = B_\perp C$ as in (19), then $\psi_\theta(u) = C$, so (90) is equivalent to (86). From (90) and (91), by an application of the chain rule

$$\mathrm{D}\Phi_\theta(u)[v] = B_\perp \mathrm{D}\varphi(C)[B_\perp^\top v]$$

for $v \in T_\theta\Theta$, where $C = \psi_\theta(u)$. Now, to show that $\mathrm{D}\Phi_\theta(u)$ is not bounded, it is enough to show that $\mathrm{D}\varphi(C)$ is not bounded. However,

$$\mathrm{D}\varphi(C)[w] = \mathrm{D}V[w]\,(\arctan(a))\,U^\top + V\mathrm{D}\,(\arctan(a))[w]\,U^\top + V\,(\arctan(a))\,\mathrm{D}U[w]$$

for $w \in \mathbb{R}^{(d-r)\times r}$, where $\mathrm{D}V$, $\mathrm{D}\,(\arctan(a))$ and $\mathrm{D}U$ denote the derivatives of V, $\arctan(a)$ and U, as functions of C, by an abuse of notation. To simplify the proof, assume, without loss of generality, that C is a square matrix (for example, if $d - r \geq r$, it is enough to add zero columns to C). With this assumption, the following formulae hold [39],

$$\mathrm{D}V[w] = V\left[F \odot \left(V^\top w\,U\,a + a\,U^\top w^\top\,V\right)\right]$$

$$\mathrm{D}\left(\arctan(a)\right)[w] = \mathrm{I}_{(d-r)}\left[V^\top w\, U\right]$$

$$\mathrm{D}U[w] = U\left[F \odot \left(a\, V^\top w\, U + U^\top w^\top\, V\, a\right)\right]$$

where F is the matrix with entries $F_{ij} = (a_j^2 - a_i^2)^{-1}$ for $i \neq j$ and $F_{ii} = 0$. However, taking $w = V\omega a U^\top$ where ω is a $(d-r)\times(d-r)$ antisymmetric matrix, yields $\mathrm{D}\left(\arctan(a)\right)[w] = 0$ and $\mathrm{D}U[w] = 0$, while

$$\mathrm{D}V[w] = V\left[G \odot \omega\right]$$

where G has matrix elements $G_{ij} = a_i^2 a_j^2/(a_j^2 - a_i^2)$ for $i \neq j$ and $G_{ii} = 0$. Clearly, these do not remain bounded as $a_i - a_j \to 0$.

Proof of (86) here, let $s = d - r$ and assume, without any loss of generality, that $s \geq r$.

Recall that, in (86), $u = B_\perp C$ where C has singular value decomposition $C = V\, a\, U^\top$. Here, V and U are orthogonal, and $a = (\alpha\,, 0_{r\times s-r})^\top$, with $r\times r$ diagonal matrix α. Write $\Phi_\theta(u)$ under the form

$$\Phi_\theta(u) = B_\perp C(u) \text{ where } C(u) = V\arctan(a)\, U^\top \tag{92}$$

Using (85), it follows

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) \;=\; \left[Q\,\exp\begin{pmatrix} 0 & -C(u)^\top \\ C(u) & 0 \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ 0_{s\times r}\end{pmatrix}\right] \tag{93}$$

where $Q = (B\,, B_\perp)$. The aim is to show this is equal to $\mathrm{Ret}_\theta(u)$, given by (21). Using the expression of $C(u)$ in (92), and performing the matrix multiplication, it is possible to check that

$$\begin{pmatrix} 0 & -C(u)^\top \\ C(u) & 0 \end{pmatrix} = \begin{pmatrix} U & \\ & V \end{pmatrix}\begin{pmatrix} 0 & -\arctan(a)^\top \\ \arctan(a) & 0 \end{pmatrix}\begin{pmatrix} U^\top & \\ & V^\top \end{pmatrix} \tag{94}$$

Recall $\exp(AXA^{-1}) = A\exp(X)A^{-1}$ for any square matrices A and X, where A is invertible. It follows from (94),

$$\exp\begin{pmatrix} 0 & -C(u)^\top \\ C(u) & 0 \end{pmatrix} = \begin{pmatrix} U & \\ & V \end{pmatrix}\exp\begin{pmatrix} 0 & -\arctan(a)^\top \\ \arctan(a) & 0 \end{pmatrix}\begin{pmatrix} U^\top & \\ & V^\top \end{pmatrix}$$

By plugging this into (93), and noticing that

$$\left[\begin{pmatrix} U^\top & \\ & V^\top \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ 0_{s\times r}\end{pmatrix}\right] = \left[\begin{pmatrix} \mathrm{I}_r \\ 0_{s\times r}\end{pmatrix}\right]$$

it follows

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) \;=\; \left[Q\begin{pmatrix} U & \\ & V \end{pmatrix}\exp\begin{pmatrix} 0 & -\arctan(a)^\top \\ \arctan(a) & 0 \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ 0_{s\times r}\end{pmatrix}\right] \tag{95}$$

If $f = (\phi\,, 0_{r\times s-r})^\top$ where ϕ is $r\times r$ diagonal, then, under the assumption that $s \geq r$, (this is proven in detail, at the end of the present proof),

$$\exp\begin{pmatrix} 0 & -f^\top \\ f & 0 \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ 0_{s\times r} \end{pmatrix} = \begin{pmatrix} C(\phi) \\ S(f) \end{pmatrix} \tag{96}$$

where $C(\phi) = \cos(\phi)$ and $S(f) = \sin(f)$, with the functions cos and sin applied to each matrix element of ϕ and f, respectively. The identity (96) can be used to evaluate the matrix exponential in (95), since $a = (\alpha\,, 0_{r\times s-r})^\top$. This yields,

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) = \left[Q\begin{pmatrix} U & \\ & V \end{pmatrix}\begin{pmatrix} \cos(\arctan(\alpha)) \\ \sin(\arctan(a)) \end{pmatrix}\right]$$

However, since $\cos(\arctan(x)) = 1/(1+x^2)^{1/2}$ and $\sin(\arctan(x)) = x/(1+x^2)^{1/2}$, this becomes

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) = \left[Q\begin{pmatrix} U & \\ & V \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ a \end{pmatrix}(\mathrm{I}_r + \alpha)^{-1/2}\right] = \left[Q\begin{pmatrix} U & \\ & V \end{pmatrix}\begin{pmatrix} \mathrm{I}_r \\ a \end{pmatrix}\right] \tag{97}$$

where the second equality holds because $(\mathrm{I}_r + \alpha)$ is invertible (the diagonal elements of α are the singular values of C, and are therefore positive). It follows from (97) that

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) = \left[Q\begin{pmatrix} U \\ Va \end{pmatrix}\right] = \left[Q\begin{pmatrix} U \\ Va \end{pmatrix}U^\top\right] \tag{98}$$

where the second equality holds because $U^\top$ is an invertible $r\times r$ matrix (which therefore does not change the span of the columns of the overall matrix product). Performing the matrix product in (98), and noting $UU^\top = \mathrm{I_r}$ and $C = V a\, U^\top$, it finally follows that

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) = \left[Q\begin{pmatrix} \mathrm{I}_r \\ C \end{pmatrix}\right] = \mathrm{Exp}_\theta(\Phi_\theta(u)) = \left[(B\,, B_\perp)\begin{pmatrix} \mathrm{I}_r \\ C \end{pmatrix}\right]$$

From $u = B_\perp C$, this immediately implies

$$\mathrm{Exp}_\theta(\Phi_\theta(u)) = [B + u]$$

which means $\mathrm{Exp}_\theta(\Phi_\theta(u)) = \mathrm{Ret}_\theta(u)$, as required in (86).

Proof of (96) this follows from

$$\exp\begin{pmatrix} 0 & -f^\top \\ f & 0 \end{pmatrix} = \begin{pmatrix} C(\phi) & -S(f)^\top \\ S(f) & C(\phi) \end{pmatrix} \tag{99}$$

where $C(\phi)$ and $S(f)$ are as in (96) and where (recall it is assumed that $s \geq r$),

$$C(\phi) = \begin{pmatrix} \cos(\phi) & 0_{r\times s-r} \\ 0_{s-r\times r} & 0_{s-r\times s-r} \end{pmatrix}$$

To prove (99), write

$$\begin{pmatrix} 0 & -f^\top \\ f & 0 \end{pmatrix} = \sum_{i=1}^{r} \phi_i \, \mathrm{b}_{r+i} \tag{100}$$

where $f = (\phi \, , 0_{r\times s-r})^\top$ with diagonal ϕ, and where

$$\mathrm{b}_{r+i} = e_{r+i,i} - e_{i,r+i}$$

with $e_{j,k}$ a matrix all of whose elements are zero, except the one at row j and column k, which is equal to 1. One easily checks the matrices b_{r+i} commute with each other. Therefore, (100) implies

$$\exp\begin{pmatrix} 0 & -f^\top \\ f & 0 \end{pmatrix} = \prod_{i=1}^{r} \exp(\phi_i \, \mathrm{b}_{r+i}) \tag{101}$$

However, it is elementary that

$$\exp(\phi_i \, \mathrm{b}_{r+i}) \; = \; \mathrm{I}_d + (\cos(\phi_i) - 1) \, \mathrm{a}_{r+i} \; + \; \sin(\phi_i) \, \mathrm{b}_{r+i} \tag{102}$$

where I_d is the $d \times d$ identity matrix and

$$\mathrm{a}_{r+i} = e_{i,i} + e_{r+i,r+i}$$

Finally, (99) follows from (101) and (102), after noting the matrix products, for $i \neq k$,

$$\begin{array}{ll} \mathrm{a}_{r+i}\mathrm{a}_{r+k} = 0 & \quad \mathrm{a}_{r+i}\mathrm{b}_{r+k} = 0 \\ \mathrm{b}_{r+i}\mathrm{a}_{r+k} = 0 & \quad \mathrm{b}_{r+i}\mathrm{b}_{r+k} = 0 \end{array}$$

which can be checked immediately.

E Proofs of Section 5.3

We present lemmas to verify **A**2–**A**3, **MD**1 for the robust barycenter problem. In addition, we show that the function $V(\theta)$ of (26) is strictly g-convex. Thus the the robust barycenter of a probability distribution π on Θ exists and is unique. Consequently, as the geodesic SA scheme (28) finds a stationary point of (26), the strict g-convexity of $V(\theta)$ guarantees that such stationary point is globally optimal and is unique.

Lemma 16. *Let Θ be a Hadamard manifold with sectional curvature bounded below by $-\kappa^2$, $x \in \Theta$ and $\delta > 0$. Define for any $\theta \in \Theta$,*

$$V_2(\theta) = \rho_\Theta^2(x,\theta) \quad \text{and} \quad V_1(\theta) = \delta^2 \left[V_2(\theta)/\delta^2 + 1\right]^{1/2} - \delta^2 \; .$$

Then, V_1 is a smooth function and its Riemannian gradient is given for any $\theta \in \Theta$ by

$$\operatorname{grad} V_1(\theta) = -\operatorname{Exp}_\theta^{-1}(x) \Big/ \left[V_2(\theta)/\delta^2 + 1\right]^{1/2} \; . \tag{103}$$

Moreover, for any $\theta \in \Theta$, $v \in \mathrm{T}_\theta\Theta \setminus \{0\}$, its Hessian satisfies

$$0 < \operatorname{Hess} V_1(\theta)(v,v) \leq (1 + \delta\kappa) \|v\|_\theta^2 \; .$$

Proof. The proof relies heavily on the computation of $\operatorname{grad} V_2$ and the Hessian comparison of V_2 done in [24, Theorem 5.6.1]. Indeed, [24, Theorem 5.6.1] shows V_2 is smooth and that for any $\theta \in \Theta$,

$$\operatorname{grad} V_2(\theta) = -2\operatorname{Exp}_x^{-1}(\theta) \; .$$

Hence, (103) follows by composition since $V_1(\theta) = \varpi \circ V_2(\theta)$ where $\varpi : t \to \delta^2[t/\delta^2+1]^{1/2} - \delta^2$ for $t \in \mathbb{R}_+$.

For any $v \in \mathrm{T}_v\Theta$, recall that $\operatorname{Hess} V_1(\theta)(v,v) = \langle \nabla_v \operatorname{grad} V_1(\theta), v\rangle_\theta$, where ∇ is the Levi-Civita connection – see Appendix A.8. The product rule for the covariant derivative [18, p. 73] for a smooth function f and vector field Y on Θ gives $\nabla(fY) = \nabla f \otimes Y + f\nabla Y$. Applying this result to

$$f(\theta) = \left[V_2(\theta)/\delta^2 + 1\right]^{-1/2} \; , \quad Y(\theta) = -\operatorname{Exp}_\theta^{-1}(x) \; ,$$

and using $Y = \operatorname{grad} V_2/2$ gives

$$\begin{aligned}\operatorname{Hess} V_1(\theta) = -\operatorname{Exp}_\theta^{-1}(x) \otimes \operatorname{Exp}_\theta^{-1}(x) \Big/ \left[\delta^2 \left\{1 + V_2(\theta)/\delta^2\right\}^{3/2}\right] \\ + \operatorname{Hess} V_2(\theta) \Big/ \left[2\left\{1 + V_2(\theta)/\delta^2\right\}^{1/2}\right] \; .\end{aligned} \tag{104}$$

Let $\theta \in \Theta$ and $v \in \mathrm{T}_\theta\Theta \setminus \{0\}$. On the one hand, we have $\operatorname{Exp}_\theta^{-1}(x) \otimes \operatorname{Exp}_\theta^{-1}(x)(v,v) = \langle \operatorname{Exp}_\theta^{-1}(x), v\rangle_\theta^2$. Therefore, using Cauchy-Schwarz inequality,

$$0 \leq \operatorname{Exp}_\theta^{-1}(x) \otimes \operatorname{Exp}_\theta^{-1}(x)(v,v) \leq \left\|\operatorname{Exp}_\theta^{-1}(x)\right\|_\theta^2 \|v\|_\theta^2 = \rho_\Theta^2(\theta,x) \|v\|_\theta^2 \; , \tag{105}$$

since $\|\operatorname{Exp}_\theta^{-1}(x)\|_\theta = \rho_\Theta(\theta,x)$. On the other hand, [24, Theorem 5.6.1] implies that

$$2\|v\|_\theta^2 \leq \operatorname{Hess} V_2(\theta)(v,v) \leq 2\kappa\rho_\Theta(\theta,x) \coth\left[\kappa\rho_\Theta(\theta,x)\right] \|v\|_\theta^2 \; . \tag{106}$$

Now, combining (105)-(106) in (104), and using $t\coth(t) \le 1+t$ for $t \ge 0$, it follows that

$$\begin{aligned}\operatorname{Hess} V_1(\theta)(v,v) &\le [1+\kappa\rho_\Theta(\theta,x)] \Big/ \left\{1+V_2(\theta)/\delta^2\right\}^{1/2} \|v\|_\theta^2 \\ &\le (1+\kappa\delta)\, \|v\|_\theta^2 \ ,\end{aligned}$$

where we have used $1+V_2(\theta)/\delta^2 \ge \max(1, \rho_\Theta^2(\theta,x)/\delta^2)$. Similarly, we obtain

$$\begin{aligned}\operatorname{Hess} V_1(\theta)(v,v) \ge &- \left(\rho_\Theta^2(\theta,x)\, \|v\|_\theta^2\right) \Big/ \left(\delta^2 \left\{1+V_2(\theta)/\delta^2\right\}^{3/2}\right) + \|v\|_\theta^2 \Big/ \left\{1+V_2(\theta)/\delta^2\right\}^{1/2} \\ &> 0 \ ,\end{aligned}$$

which concludes the proof. ☐

Lemma 17. *Let Θ be a Hadamard manifold with sectional curvature bounded below by $-\kappa^2$. Furthermore, consider the Lyapunov function given by (26) and stochastic approximation scheme (28). Then,* ***A*** *2 and* ***A*** *3 are satisfied, with $\underline{c} = \overline{c} = 1$, $L = 1+\delta\kappa$, and $b_\infty = 0$. Moreover,* ***MD*** *1 is satisfied, with $\sigma_0 = \delta^2$ and $\sigma_1 = 0$.*

Proof. To show that **A** 2 holds with the stated values of $\underline{c}$, $\overline{c}$ and L, note that the scheme (28) can be written

$$\theta_{n+1} \ = \ \operatorname{Exp}_{\theta_n}\left(\eta_{n+1}\, H_{\theta_n}(X_{n+1})\right) \ ,$$

where the stochastic update $H_{\theta_n}(X_{n+1})$ is given by

$$H_\theta(x) \ = \ -\operatorname{Exp}_\theta^{-1}(x) \Big/ \left[1+\{\rho_\Theta(\theta,x)/\delta\}^2\right]^{1/2} \ . \tag{107}$$

Then, both $\operatorname{grad} V$ and $\operatorname{Hess} V$ can be computed differentiating under the integral in (26). Using Lemma 16, we know that for any $x \in \Theta, V_1 : \theta \mapsto \delta^2(1+\rho_\Theta^2(\theta,x)/\delta^2)^{1/2} - \delta^2$ is smooth. Using $\|\operatorname{Exp}_\theta^{-1}(x)\|_\theta = \rho_\Theta(\theta,x)$, we have that for any $x,\theta \in \Theta$, $\|\operatorname{grad} V_1(\theta)\|_\theta \le 1$ and, for any $x,\theta \in \Theta, v \in \mathrm{T}_\theta\Theta$,

$$|\operatorname{Hess} V_1(\theta)(v,v)| \le (1+\kappa\delta)\, \|v\|_\theta^2 \ .$$

Under these domination conditions, using Lebesgue's dominated convergence theorem, we have for any $\theta \in \Theta$,

$$\operatorname{grad} V(\theta) = -\int_\Theta \operatorname{Exp}_\theta^{-1}(x) \Big/ \left[1+\{\rho_\Theta(\theta,x)/\delta\}^2\right]^{1/2} \pi(\mathrm{d}x) \ ,$$

and for any $v \in \mathrm{T}_\theta\Theta$,

$$\operatorname{Hess} V(\theta)(v,v) = \int_\Theta \operatorname{Hess} V_1(\theta)(v,v)\pi(\mathrm{d}x) \ .$$

Thus, for any $v \in \mathrm{T}_\theta\Theta \setminus \{0\}$,

$$0 < \operatorname{Hess} V(\theta)(v,v) \leq (1+\kappa\delta)\, \|v\|_\theta^2 \ . \tag{108}$$

This last inequality proves that the operator norm of Hess V is upper bounded by $1+\delta\kappa$. Therefore, using Lemma 10, it follows that **A**2-(b) holds with $L = 1+\delta\kappa$.

It remains to prove that Assumptions **MD**1 is satisfied with $\sigma_0 = \delta^2$ and $\sigma_1 = 0$. To do so, note first from (107), that

$$\|H_{\theta_n}(X_{n+1})\|_{\theta_n}^2 = \left\|\mathrm{Exp}_{\theta_n}^{-1}(X_{n+1})\right\|_{\theta_n}^2 \Big/ \left(1 + \{\rho_\Theta(\theta, X_{n+1})/\delta\}^2\right) \ .$$

Since $\|\mathrm{Exp}_{\theta_n}^{-1}(X_{n+1})\|_{\theta_n}^2 = \rho_\Theta^2(\theta_n\,, X_{n+1})$, it follows that $\|H_{\theta_n}(X_{n+1})\|_{\theta_n}^2 \leq \delta^2$. Thus, in the notation of (5),

$$\mathbb{E}\left[\|e_{\theta_n}(X_{n+1})\|_{\theta_n}^2 \middle| \mathcal{F}_n\right] = \mathbb{E}\left[\|H_{\theta_n}(X_{n+1}) - h(\theta_n)\|_{\theta_n}^2 \middle| \mathcal{F}_n\right] \leq \delta^2 \tag{109}$$

since the conditional variance is bounded by the mean square. Now, the required values of σ_0 and σ_1 can be read from (109). □

Proposition 18. *Let Θ be a Hadamard manifold with sectional curvature bounded below, and let π be a probability distribution on Θ. Assume there exists some $\tau \in \Theta$ such that*

$$\int_\Theta \rho_\Theta(\tau, x)\, \pi(\mathrm{d}x) \ < \ +\infty \ . \tag{110}$$

Then, the function $V : \Theta \to \mathbb{R}_+$ given by (26) is geodesically strictly convex, but not strongly convex, in general. Moreover, V has a unique global minimizer $\theta^* \in \Theta$. In other words, π has a unique robust barycenter θ^*.

Proof. Under condition (110), the function V takes finite values, $V(\theta) < +\infty$ for any $\theta \in \Theta$. Indeed, note the following inequality, which holds for all $x \geq 0$,

$$\delta^2 \left[1 + \{x/\delta\}^2\right]^{1/2} - \delta^2 \ \leq \ \delta\, x$$

From (27) and (26), this inequality implies

$$V(\theta) \ \leq \ \delta \int_\Theta \rho_\Theta(\theta, x)\, \pi(\mathrm{d}x) \qquad \text{for } \theta \in \Theta$$

and, furthermore, by the triangle inequality,

$$V(\theta) \ \leq \ \delta \int_\Theta \left(\rho_\Theta(\theta,\tau) + \rho_\Theta(\tau,x)\right) \pi(\mathrm{d}x) \ = \ \rho_\Theta(\theta,\tau) + \int_\Theta \rho_\Theta(\tau,x)\, \pi(\mathrm{d}x) \tag{111}$$

Then, it follows from (110) and (111) that $V(\theta) < +\infty$ for $\theta \in \Theta$.

Further, from (108) in the proof of Lemma 17, $\operatorname{Hess} V(\theta) \succ 0$, so V has strictly positive-definite Riemannian Hessian, and is therefore geodesically strictly convex. To see that $V(\theta)$ may fail to be strongly convex, consider the case where $\pi = \delta_\tau$ (here, δ_τ is the Dirac distribution, concentrated at $\tau \in \Theta$). By (26), it then follows

$$V(\theta) \;=\; \tilde{\rho}(\theta,\tau) \tag{112}$$

Then, let $\gamma(t)$ be a geodesic through τ, given by $\gamma(t) = \operatorname{Exp}_\tau(t\,u)$ for $t \in \mathbb{R}$, where $u \in T_\tau\Theta$ has $\|u\|_\tau = 1$. If $V(\theta)$ is given by (112), then it follows from (27)

$$(V \circ \gamma)(t) \;=\; \delta^2\left[1 + \{t/\delta\}^2\right]^{1/2} \,-\, \delta^2$$

but this is not a strongly convex function of $t \in \mathbb{R}$. Therefore, V is not geodesically strongly convex on Θ [41, p. 187].

It remains to show that V has a unique global minimizer θ^*. Since V is geodesically strictly convex, and bounded below (indeed, V is positive), it is enough to show that V is coercive, in the sense that $V(\theta) \to +\infty$ when $\rho_\Theta(\theta,\tau) \to +\infty$. To do so, note the following inequality holds for all real x,

$$\delta^2\left[1 + \{x/\delta\}^2\right]^{1/2} \,-\, \delta^2 \;\geq\; \frac{\delta}{\sqrt{2}}(x - \delta)$$

Then, using (27) and (26), this inequality implies

$$V(\theta) \;\geq\; \frac{\delta}{\sqrt{2}}\int_\Theta \left(\rho_\Theta(\theta,x) - \delta\right)\,\pi(\mathrm{d}x)$$

or, by the triangle inequality,

$$V(\theta) \;\geq\; \frac{\delta}{\sqrt{2}}\int_\Theta \left(|\rho_\Theta(\theta,\tau) - \rho_\Theta(\tau,x)| - \delta\right)\,\pi(\mathrm{d}x) \qquad \text{for } \theta \in \Theta$$

However, this directly yields

$$V(\theta) \;\geq\; \frac{\delta}{\sqrt{2}}\,\rho_\Theta(\theta,\tau) - \frac{\delta}{\sqrt{2}}\int_\Theta \left(\rho_\Theta(\tau,x) + \delta\right)\,\pi(\mathrm{d}x)$$

Clearly, the right-hand side increases to $+\infty$ when $\rho_\Theta(\theta,\tau) \to +\infty$. Thus, V is indeed coercive.

Finally, to show that V has a unique global minimizer θ^*, let $V_* = \inf\{V(\theta)\,;\theta \in \Theta\}$ and $(\theta_n\,;n = 1,2,\ldots)$ a sequence of points in Θ such that $\lim V(\theta_n) = V_*$. since V takes finite values, $V_* < +\infty$. Therefore, there exists some $R > 0$ such that $\rho_\Theta(\theta_n\,,\tau) < R$ for all n. This is because, otherwise, $\rho_\Theta(\theta_n\,,\tau) \to +\infty$, and thus $V(\theta_n) \to +\infty$, since V is coercive. Because Θ is a complete Riemannian manifold, the metric ball $B(\tau,R)$ has compact closure (a consequence of the Hopf-Rinow theorem [28]). This implies that the sequence θ_n has a convergent subsequence, whose limit θ^* belongs to the closure of $B(\tau,R)$. By continuity of V, it is clear that $V(\theta^*) = \lim V(\theta_n) = V_*$, so θ^* is indeed a global minimizer of V. This global minimizer is unique because V is geodesically strictly convex. $\square$